UNIVERSITÀ DEGLI STUDI DELL'AQUILA

DIPARTIMENTO DI SCIENZE FISICHE E CHIMICHE

Dottorato di Ricerca in Scienze Fisiche e Chimiche

XXXIV ciclo

Traceability and Authenticity of high-valued dairy products by combining Analytical and Chemometric methods

SSD CHIM/01

Francesca Di Donato

Coordinatore del corso

Prof. Massimiliano Aschi

Tutor

Prof. Angelo Antonio D'Archivio

Dr. Alessandra Biancolillo

A.A. 2020/2021

To my grandparents,
the brightest stars in the sky

To my uncle,
to his strength,
to his long struggle

Acknowledgements

People are forced to leave their home and my research can't save any of them.

Nevertheless, it is still full of meaning for me, even in a historical period like the one we are experiencing. So, I wish to acknowledge everyone who made my work possible.

I am deeply grateful to my supervisors Prof. Angelo Antonio D'Archivio and Dr. Alessandra Biancolillo. They have assisted and guided me at every step of my path, offering wise advices and chances to grow, both personally and professionally.

I would like to extend my heartfelt thanks to Prof. Fabrizio Ruggieri and Dr. Maria Anna Maggi. They were the first to take me in.

My gratitude goes also to Prof. Rasmus Bro who hosting me in Copenhagen during my period abroad as visiting PhD student. He has been for me a great mentor just as much as a great human being. I will never forget his kindness, his generosity, his "fingers crossed" during our meetings, and his "it's almost Friday" on Tuesday.

I would like to acknowledge all the researchers I collaborate with for the realization of this thesis and more: Prof. Federico Marini from La Sapienza University, Prof. Nicoletta Spreti and Cinzia Casieri, and Dr. Francesco Gabriele, Prof. Armando Carlone, Dr. Valeria Nori and Arianna Sinibaldi, and Dr. Martina Foschi from L'Aquila University. They led me to face a new way of thinking and to experience the real essence of the art of *collaboration*.

I would like to thank all my student interns, my former and current colleagues, all the people I met in these years, and all my friends. They shared with me passion, joy, struggles, laughter, and life.

Today, some of them are bits of me.

A last thank to my family and Federico for their unwavering support.

I wouldn't be me without you.

The work described in this thesis was carried out at:

DSFC-Dipartimento di Scienze Fisiche e Chimiche
Università degli Studi dell'Aquila
L'Aquila, Italy

Dipartimento di Chimica
Università La Sapienza
Roma, Italy

Abstract

An extended and updated list of analytical techniques for confirmation of food authentication and to support law enforcement is nowadays strongly demanded by food safety and quality control organizations all over the world, presiding over the farm to fork continuum of each specific sector of food.

Analytical Chemistry can give a decisive answer in this direction playing an active role in restoring dignity to traditional food species, due to the potentiality of many available techniques to provide traceability and authenticity information. The scientific community is indeed offering proven, accurate and reliable methods to protect and clearly distinguish high-valued food products from any illegal substitutes setting-up applications as fast, efficient, cheap, and easily applicable (at custom level) as possible.

Into an era of high-throughput analytical devices, Chemometrics, giving the necessary multivariate statistical-mathematical and computing expertise, is co-evolving fast together with Analytical Chemistry. From the food-traceability and authentication point of view, in particular, Chemometrics grants the need to extract only relevant information and obtain reliable models capable of inferring the underlying relationships which link the complex compositional profile of food products and processing conditions to very general end-properties of foodstuffs, such as healthiness, consumer perception, or to a specific territory. Authenticity and traceability, then, are inherently multivariate, covering many different aspects: chemical and physical characterization, adulteration, discrimination, mislabeling, and monitoring of the production process.

Under these perspectives, the research studies the present PhD thesis lays on involve the development of new analytical methods reinforced by Chemometrics to handle different food traceability and authentication cases. The bulls-eye has been focused on the dairy sector, characterized by several decades of increased market disturbance. A number of intervention mechanisms have been indeed put in place to promote and protect the milk sector and give it back value. Nowadays, milk sector represents a significant proportion of the value of European Union-agricultural output. Total EU milk production is estimated to be second only to India; moreover, EU is a major exporter of dairy products and the biggest one in the world for cheese.

More in details, the reader will be led through different applications handling several food-related issues, focusing on dairy products: from the authentication of cheeses labeled with quality schemes, to the contamination of niche products with cheaper one, and the monitoring of storage conditions. Both *targeted* analysis, giving information about singular molecular and atomic species, and fingerprinting techniques, providing *convoluted* profiles of foodstuffs, have been employed and compared in performances in this thesis. Exploratory multivariate analysis and different single- and multi-block supervised pattern recognition methods have been employed to interpret the instrumental outputs. The excellent results obtained in the reported applications support

the effective and indispensable role of chemometric processing in the field of food control to handle such complex matrices, like dairy products are.

For my own part I hope to have made a good service to the scientific community, providing another *brick in the wall* in such greatly discussed and of real-time-interest field of knowledge.

Publications

This PhD thesis has been partially published in the following papers:

Di Donato, F.; Foschi, M.; Vlad, N.; Biancolillo, A.; D'Archivio, A.A. *Multi-elemental composition data handled by Chemometrics for the discrimination of high-value Italian Pecorino Cheeses. Molecules,* **2021**, *26(22)*, 6875. https://doi.org/10.3390/molecules26226875

Di Donato, F.; Biancolillo, A.; Mazzulli, D.; Rossi, L.; D'Archivio A.A. *HS-SPME/GC–MS volatile fraction determination and chemometrics for the discrimination of typical Italian Pecorino cheeses. Microchemical Journal*, **2021**, *165*, 106133. https://doi.org/10.1016/j.microc.2021.106133

Aït-Kaddour, A.; Hassoun, A.; Bord, C.; Schmidt-Filgueras, R.; Biancolillo, A.; **Di Donato, F.**; Tümay Temiz, H.; Cozzolino, D. *Application of Spectroscopic Techniques to Evaluate Heat Treatments in Milk and Dairy Products: an Overview of the Last Decade. Food and Bioprocess Technology*, **2021**, *14*, 781-803. https://doi.org/10.1007/s11947-021-02607-0

Di Donato, F.; Biancolillo, A.; Foschi, M.; D'Archivio, A.A. *Application of SPORT algorithm on ATR-FTIR data: a rapid and green tool for the characterization and discrimination of three typical Italian Pecorino cheeses (Submitted to Journal of Food Composition and Analysis).*

Di Donato, F.; Biancolillo, A.; Ferretti, A.; D'Archivio, A.A.; Marini, F. *Near Infrared Spectroscopy coupled to chemometrics for the authentication of donkey milk. Submitted to Food Control.*

Table of Contents

Chapter 1: Food Traceability and Authentication: an overview **1**

1.1 The food-related international panorama 1

1.1.1 Food safety and quality-the response of the governments 1

1.1.2 Food laws and regulations-Food control system 6

1.1.3 Food control system-the response of European Union 9

1.2 Food Traceability 11

1.2.1 The general framework of Food Traceability 11

1.2.2 Does Traceability lead to Food Authentication? 18

1.2.3 Quality indicators-the international and national panorama 21

1.3 Focus on dairy sector 26

1.3.1 EU regulations on milk and dairy products 26

1.3.2 European Milk Market Observatory 28

1.4 Aim and Outline of the work 31

Abbreviations 35

Chapter 2: Analytical Chemistry: a force for food safety and quality **37**

2.1 Introduction 37

2.2 Experimental methods 40

2.2.1 Gas-Chromatography (GC) 40

2.2.2 Multi-elemental profile-Inductively Coupled Plasma-Optical Emission Spectrometry (ICP-OES) 44

2.2.3 Fourier Transform-InfraRed (FT-IR) Spectroscopy 48

2.2.3.1 Near InfraRed (NIR) Spectroscopy 48

2.2.3.2 Mid InfraRed (MIR) Spectroscopy 50

2.2.4 Two-Dimensional Proton Nuclear Magnetic Resonance Relaxometry (2D 1H NMR-R) 52

Abbreviations 56

Chapter 3: Someone could ask: why Chemometrics? **57**

3.1 Introduction 57

3.2 Exploratory Data Analysis (EDA) 62

3.3 Coupling of EDA and experimental design: from ANOVA to ASCA 66

3.3.1 ANalysis Of VAriance (ANOVA) 67

3.3.2 ANOVA applied to designed experiments 68
3.3.3 ANOVA-Simultaneous Component Analysis (ASCA) 71
3.4 Predictive models 73
3.4.1 Classification methods 74
3.4.2 Discriminant classification: Linear Discriminant Analysis (LDA) and Partial Least Squares-Discriminant Analysis (PLS-DA) 75
3.4.3 Multi-block (MB) data analysis: an overview 80
Abbreviations 88
Section 1: Multi-methodological characterization coupled with Chemometrics in the discrimination of typical Pecorino cheeses from Central Italy 89
S1.1 Introduction 91
S1.2 Aim and Outline of the project 93
S1.2.1 Why the volatile profile? 97
S1.2.1 Why the multi-elemental composition? 98
S1.2.1 Why the fatty acids profile? 99
S1.2.1 Why the infrared-spectroscopy *fingerprinting*? 100
Chapter 4 103
4.1 Materials and Methods 103
4.1.1 Pecorino cheese samples and chemicals 103
4.1.2 Headspace solid-phase micro-extraction by design of the experiments (DOE) and Response Surface Methodology (RSM) 104
4.1.3 Gas chromatography-mass spectrometry (GC-MS) analysis 104
4.1.4 Statistical Analysis 105
4.1.4.1 Optimization of HS-SPME extraction by design of the experiments (DOE) and Response Surface Methodology (RSM) 105
4.1.4.2 Discrimination of Pecorino cheeses 106
4.2 Results and Discussion 106
4.2.1 Investigation by DOE-RSME of the SPME efficiency 106
4.2.2 HS-SPME/GC-MS profile of Pecorino cheese aroma 109
4.2.2.1 Acids 112
4.2.2.2 Carbonyl compounds 112
4.2.2.3 Esters 113
4.2.2.4 Alcohols 113
4.2.2.1 Others 114
4.2.1 PCA exploration of GC data 114

4.2.4 Discrimination of Pecorino cheese samples 116
4.2.4.1 LDA model 116
4.2.4.2 PLS-DA model 118
4.3 Conclusions 121
Chapter 5 ..... **123**
5.1 Materials and Methods 123
5.1.1 Pecorino cheese samples and chemicals 123
5.1.2 Sample preparation and Microwave-Assisted Digestion 124
5.1.3 ICP-OES Analysis 124
5.1.4 Statistical Analysis 125
5.2 Results and Discussion 125
5.2.1 Validation of ICP-OES Analysis of Pecorino cheeses 125
5.2.2 Exploratory Analysis of ICP-OES data 127
5.2.3 PLS-DA geographical classification of Pecorino 130
5.3 Conclusions 131
Chapter 6 ..... **133**
6.1 Materials and Methods 133
6.1.1 Pecorino cheese samples and chemicals 133
6.1.2 Fat extraction 134
6.1.2.1 Triacylglycerols (TAGs) 134
6.1.2.2 Free Fatty Acids (FFAs) 135
6.1.3 GC-FID Analysis 135
6.1.4 Statistical Analysis 136
6.2 Results and Discussion 136
6.2.1 Fat composition 136
6.2.2 Discrimination of Pecorino cheese samples 141
6.2.2.1 PLS-DA Analysis 142
6.2.2.2 MB Analysis 142
6.3 Conclusions 144
Chapter 7 ..... **145**
7.1 Materials and Methods 145
7.1.1 Pecorino cheese samples 145
7.1.2 ATR FT-IR measurements 145
7.1.3 Statistical Analysis 146

7.2 Results and Discussion 146
7.2.1 ATR FT-IR Characterization 146
7.2.2 Discrimination of Pecorino cheese samples 149
7.2.2.1 PLS-DA Analysis 150
7.2.2.2 SPORT Analysis 151
7.3 Conclusions 154
Chapter 8 ..... **155**
8.1 Materials and Methods 155
8.1.1 Pecorino cheese samples 155
8.1.2 Statistical Analysis 156
8.2 Results and Discussion 156
8.2.1 Discrimination of Pecorino cheese samples 156
8.2.1.1 PLS-DA Analysis 158
8.2.1.2 SO-PLS-LDA Analysis 158
8.3 Conclusions 160
Section 2: Green approaches for the characterization inline of dairy commodities **161**
Chapter 9 ..... **163**
9.1 Introduction 163
9.2 Materials and Methods 165
9.2.1 Pecorino cheese samples 165
9.2.2 2D ^{1}H NMR-R measurements 165
9.2.3 Statistical Analysis 166
9.3 Results and Discussion 167
9.3.1 2D ^{1}H NMR-R measurements and processing 167
9.3.2 Selection of the smoothing coefficient α 169
9.3.3 ASCA Analysis 173
9.4 Conclusions and perspectives 175
Chapter 10 ..... **177**
10.1 Introduction 177
10.2 Materials and Methods 179
10.2.1 Donkey milk samples 179
10.2.2 NIR measurements 180
10.2.3 Statistical Analysis 180
10.3 Results and Discussion 181

10.3.1 Authentication of donkey milk: PLS-DA classification of milk samples 182
10.3.2 Quantification of cow milk content in donkey milk samples 186
10.3.3 ASCA Analysis 187
10.3.3.1 Estimation of the effect and its significance 187
10.3.3.2 SCA Models 188
10.4 Conclusions 190
Final Remarks **191**
Appendices **193**
Appendix A: Data pre-treatment **195**
Appendix B: Validation **203**

Chapter 1

Food Traceability and Authentication: an overview

1.1 The food-related international panorama

1.1.1 Food safety and quality-the response of the governments

A national food control system ensures that food available within a Country conforms to food safety and quality requirements and that is honestly and accurately labeled as prescribed by the law. As such, food control systems protect the health and safety of consumers and help assure the safety and quality of foods being traded both nationally and internationally.

The terms *food safety* and *food quality* refer to two slightly different concepts so they could sometimes be confused. *Food safety* refers to the conditions and practices that preserve the quality of the food to prevent all those hazards, either chronic or acute, which make food dangerous for human consumption causing foodborne illness. Factors which contribute to potential hazards in foods include improper agricultural practices, poor hygiene at all stages of the food chain, lack of preventive controls in food processing and preparation operations, misuse of chemicals, contaminated raw materials, ingredients, and water, inadequate or improper storage, etc. The term *food quality* includes all the negative (spoilage, contamination, discoloration, off-odors) and positive (origin, color, flavor, texture, processing method of the food) attributes that could influence a product's value to the consumer. This distinction between safety and quality has implications in public policy and influences the nature and content of the food control system most suited to meet predetermined national objectives.

With an estimated 600 million cases[1] of foodborne illnesses and 420000 deaths[2] annually, unsafe food is a threat to human health and economies globally. Therefore, ensuring food safety is a public health priority and an essential step to achieve food security. In an age of rapidly changing food technologies and ever-increasing global food trade, controlling foodborne hazards along the entire food supply-chain, that nowadays cross multiple national borders, has taken on a pivotal role. In order to cope with complex emerging

---

[1] Food and Agriculture Organization of the United Nations. Food safety and quality. Retrieved from https://www.fao.org/food-safety/background/en/ (last access January 2021).

[2] World Health Organization. Food safety. Retrieved from https://www.who.int/news-room/fact-sheets/detail/food-safety (last access January 2022).

challenges of 21st century, food control systems need to be up to date with the latest developments and be harmonized between different governments, producers and consumers. In this scenario, all the national and international organizations had put in place different policies drawing attention to the need of establishing general principles and requirements concerning food and feed law.

For over 50 years, Food and Agriculture Organization of the United Nations (FAO) in collaboration with the World Health Organization (WHO) has been providing neutral and independent scientific advice as the basis for the international food safety standards and guidelines established by the Codex Alimentarius Commission[3] (CAC). More in details, CAC establishes standards for food commodities and particular regulations about many kinds of contaminants, labeling, analytical procedures, import/export inspections and certifications, and other parameters and practices. All the Nations members of the CAC (all EU Countries are) should exploit Codex Alimentarius as a benchmark to establish policies considering each national food control apparatus.

In order to operate effectively, a food control system requires appropriate legal and policy instruments, well-qualified human resources, sound institutional frameworks as well as financial assets, equipment, and infrastructures (including access to laboratories). In this perspective, FAO and WHO are working through the development of a framework collecting different joint FAO/WHO expert bodies and meetings, such as the Expert Committee on Food Additives[4] (JECFA, active since 1956) and the Expert Meetings on Microbiological Risk Assessment[5] (JEMRA, active since 2000) to name a few.

Moreover, in 2002, European Union (EU) food law was revamped following a series of food crises (such as the Bovine Spongiform Encephalopathy (BSE), or mad cow disease, dioxins in food and feed, Foot and Mouth Disease (FMD), outbreaks of foodborne illnesses) in the late 1990s that shook consumers' trust in food, resulting in disruptions of trade and questioning the EU's ability to manage food risks. The European Parliament reacted with the adoption of the Regulation (European Commission, EC) No 178/2002[6], laying down the General Food Law Regulation (GFL). GFL is now the foundation of food and feed principles, requirements and procedures and the cornerstone of the EU regulatory framework since it covers all sectors of the entire supply-chain, including feed

---

[3] Food and Agriculture Organization of the United Nations & World Health Organization. Codex Alimentarius-International Food Standards. Retrieved from https://www.fao.org/fao-who-codexalimentarius/en/ (last access January 2022).

[4] Food and Agriculture Organization of the United Nations & World Health Organization. Joint Expert Committee on Food Additives. Retrieved from https://www.fao.org/food-safety/scientific-advice/jecfa/en/ (last access January 2022).

[5] Food and Agriculture Organization of the United Nations & World Health Organization. Joint Expert Meetings on Microbiological Risk Assessment. Retrieved from https://www.fao.org/food-safety/scientific-advice/microbiological-risks-and-jemra/en/ (last access January 2022).

[6] European Parliament and Council of the European Union. Regulation (EC) No 178/2002. **2002** Retrieved from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32002R0178 (last access January 2022).

production, primary production, food processing, storage, transport, and retail sale. This integrated approach to food safety, the so-called *from farm to fork*, ensures a high level of protection of human life and the effective functioning of the internal market, setting out an overarching and coherent framework for the development of food and feed legislation both at Union and National levels. The *from farm to fork* program (depicted in Figure 1.1) establishes the national enforcement of the authorities performing the controls on farms, the border controls for imported animals, plants, and specific foodstuffs, the controls during the transportation within the EU, and the controls in supermarkets, retailers, and restaurants. All these controls check for: hormones, chemical residues, bacterial/viral contamination, overall hygiene, proper refrigeration, animal and plant health requirements, animal welfare, labeling, and fraud.

Figure 1.1. *From Farm to Fork*: Controlling the safety of agri-food chain–infographic. Reproduced from https://ec.europa.eu/food/horizontal-topics/communication-outlets/infographics-factsheets_en (last access January 2022).

More in details, GFL Regulation establishes:

- the institution in 2003 of the European Food Safety Authority[7] (EFSA) as independent European agency entrusted with EU risk assessment. EFSA is funded by the EU budget, but operates separately from the EC, the European Parliament, and the EU Countries.

[7] European Commission. European Food Safety Authority (EFSA)-Trusted Science for Safe Food. Retrieved from https://www.efsa.europa.eu/en (last access January 2022).

- the institution of tools and procedures for the prevention and management of food crises, such as the Rapid Alert System for Food and Feed[8] (RASFF), a network for the swift exchange of information relating to food and feed risks. In the context of globalization, the RASFF contributed to ensure a rapid access to information for food and feed for 113 Countries outside the EU and the European Free Trade Association[9] (EFTA).

- the common definitions (such as 'food' and 'food/feed business operator'), objectives, and general principles to underpin national and EU food law. The most important principle is the assumption that the food law must be science-based (risk analysis principle). When preparing its policy and proposals relating to food safety, indeed, the EC relies on two independent Scientific Committees: Scientific Committee on Consumer Safety[10] (SCCS) and Scientific Committee on Health, Environment and Emerging Risks [11] (SCHEER). Other important principles are the public consultation throughout the decision-making cycle, the protection of consumers' interests, and the right of the public to be informed where there are reasonable grounds to suspect that a food or feed may present a risk for human or animal health.

- the general requirements addressed to food and feed business operators relating to controls for checking compliance with EU and national food law, food and feed safety, traceability, and withdrawals/recalls of unsafe food and feed.

The GFL Regulation has deeply contributed to improve food and feed traceability in the entire agri-food chain, to clarify responsibilities of the business operators, to place on the market safe food and ensure compliance with EU requirements, and to increase the international competitiveness and reputation of the EU food and drink industry for its production of high quality products (the value of food and drink sector has increased of 72% in the last decade).

EU, as a world's biggest importer and exporter of food, has been involved in cooperation with the World Trade Organization [12] (WTO, established in 1995) and the main international standard setting bodies active in the area of food safety and animal health

---

[8] European Commission. Rapid Alert System for Food and Feed (RASFF)-Food and Feed Safety Alerts. Retrieved from https://ec.europa.eu/food/safety/rasff-food-and-feed-safety-alerts_en (last access January 2022).
[9] European Free Trade Association (EFTA). Retrieved from https://www.efta.int/ (last access January 2022).
[10] European Commission. Scientific Committee on Consumer Safety (SCCS)-Public Health. Retrieved from https://ec.europa.eu/health/scientific_committees/consumer_safety_en (last access January 2022).
[11] European Commission. Scientific Committee on Health, Environmental and Emerging Risks (SCHEER-Public Health. Retrieved from https://ec.europa.eu/health/scientific_committees/scheer_en (last access January 2022).
[12] World Trade Organization. Retrieved from https://www.wto.org/ (last access January 2022).

(Codex Alimentarius, International Plant Protection Convention [13] (IPPC), World Organization for Animal Health[14] (OIE)), as certified by the stipulation of the Agreement on the Application of Sanitary and Phytosanitary Measures[15] (SPS Agreement) and the Agreement on Technical Barriers to Trade[16] (TBT). SPS Agreement states the right of WTO members to apply legal actions to guarantee humans, animals, and plants defense, of course under the constraint of exploiting scientifically accredited measures which must not cause limitations on the international commerce. TBT agreement aims to ensure that technical regulations, standards, and conformity assessment procedures are non-discriminatory and do not create unnecessary obstacles to trade, i.e., regulations on traditional quality elements, frauds, packaging and labeling determined by each Country must not be more binding on imported food with respect to local products.

The provision of safe, nutritious, high quality and affordable food to Europe's consumers is the central objective of EU policy. Its standards and requirements aim to ensure a high level of food safety and nutrition within an efficient, competitive, sustainable, and innovative global market. However, a series of emerging challenges and risks (demographic imbalances, climate change, resource and energy scarcity, slowing agricultural productivity, increasing concentration of the supply-chain, price volatility, changing diet trends and the emergence of anti-microbial resistant strands) could put the currently successful European food system under severe stress. To deal with these new challenges, the Commission has allocated 1.68 billion € to support the food safety initiative, as part of the so-called Single Market Program. The project should aim to provide insight and guidance for future policy identifying the:

- critical challenges to EU food legislative framework;
- future evolution of the challenges (in years 2020, 2030 and 2050);
- potential critical changes in the current framework necessary to maintain the prevailing high standards.

---

[13] Food and Agriculture Organization of the United Nations. International Plant Protection Convention (IPPC). Retrieved from https://www.ippc.int/en/ (last access January 2022).

[14] World Organization for Animal Health. OIE-Protecting animals, preserving our future. Retrieved from https://www.oie.int/en/home/ (last access January 2022).

[15] World Trade Organization. Agreement on the Application of Sanitary and Phytosanitary Measures. Retrieved from https://www.wto.org/english/tratop_e/sps_e/spsund_e.htm (last access January 2022).

[16] World Trade Organization. Agreement on Technical Barriers to Trade. Retrieved from https://www.wto.org/english/tratop_e/tbt_e/tbt_e.htm (last access January 2022).

1.1.2 Food laws and regulations-Food control system

Food control has been defined by FAO as "*a mandatory regulatory activity of enforcement by national or local authorities to provide consumer protection and ensure that all foods during production, handling, storage, processing, and distribution are safe, wholesome and fit for human consumption; conform to safety and quality requirements; and are honestly and accurately labelled as prescribed by law*".

Effective national food control systems are therefore essential to protect the health and safety of domestic consumers and they are also critical in enabling Countries to assure the safety and quality of their foods entering international trade and to ensure that imported foods conform to national requirements. The new global environment for food trade places considerable obligations on both importing and exporting Countries to strengthen their food control systems and to implement and enforce risk-based food control strategies. In many Countries, effective food control is undermined by the existence of fragmented legislation, multiple jurisdictions, and weaknesses in surveillance, monitoring, and enforcement. Moreover, globalization of the food supply-chain, the increasing importance of the CAC, and the obligations emerging from the WTO Agreements have resulted in unprecedented interest in the development of food standards and regulations, and the strengthening of food control infrastructure at the Country level. For this reason, international organizations, such as FAO and WHO, have prepared guidelines to enable national authorities to improve their food control systems, showing a strong interest in promoting national food control based upon scientific principles and guidelines. The publication *Assuring Food Safety and Quality: Guidelines for Strengthening National Food Control Systems*[17], that replaced the earlier FAO/WHO publication *Guidelines for Developing an Effective National Food Control System*[18], is nowadays the benchmark for developing Countries in planning, organizing, and implementing their national food control program. The term *Food Control System* is used in these guidelines to describe the integration of a mandatory regulatory approach with preventive and educational strategies that protect the whole food chain from primary producer through consumer, often described as the *farm-to-table* continuum. Consumers indeed expect protection from hazards occurring along the entire supply-chain and this protection will only occur if all sectors (farmers, industries, consumers, and so on) operate in an integrated way. An ideal food control system addresses all stages of the *farm-to-table* continuum.

---

[17] Food and Agriculture Organization of the United Nations & World Health Organization. **2003**. *Assuring Food Safety and Quality: Guidelines for Strengthening National Food Control Systems*.

[18] Food and Agriculture Organization of the United Nations & World Health Organization. **1976**. *Guidelines for Developing an Effective National Food Control System.*

While the components and priorities of a national food control system could vary from one Country to another, all food control systems typically comprise a set of five common building blocks:

- Food Law and Regulations.

  The term "food law" applies to legislation which regulates the production, trade and handling of food and hence covers the regulation of food control, food safety, quality and relevant aspects of food trade across the entire food chain, from the provision for animal feed to the consumer. The development of relevant and enforceable food laws and regulations is an essential component of a modern food control system. Nevertheless, many Countries have inadequate food legislation, and this will impact on the effectiveness of all food control activities carried out in the Country. Ideally, governments should update food legislation and food standards, harmonizing them with the CAC directives in order to provide a greater health preservation and include unambiguous definitions to increase uniformity and security.

  Historically, food legislation was simply constituted by a list of definitions of "unsafe food" and it contained all the guidelines for removing dangerous commodities from the market and on prosecuting those responsible for such violations. In the current legislation, the rules that allow the authorities introducing new propositions to reduce the food-related risks have been implemented under a more preventive and comprehensive perspective. The so-called "horizontal" food standards have nowadays replaced the older highly prescriptive ones. They presuppose a very controlled process, which must be supported by reliable data regarding the food-related risks and valid risk administration plans. Following the guidelines for the development of food laws, contained in Annex 6 of the above-mentioned joint FAO/WHO publication[17], the legislation should be transparent and based on unequivocal and high-level scientific judgments and it should include prescriptions about precautions in high risk situations where a complete risk assessment cannot be carried out. It should promote a more easily access of citizens to information concerning both the aspect of safety and traceability of commodities. Producers and processors should be definitely designated as responsible for both quality and safety and it should be ensured that only safe food with clear labelling can access the market.

- Food Control Management.

  To avoid any lack of coordination among all the partners involved in the *from farm to fork* continuum at national level, an effective food control system should include the establishment of a leadership function and administrative structures with clearly defined accountability for issues.

- Inspection Services.

  The reputation and integrity of a food control system depends, to a very large extent, on the integrity and skills of a qualified, trained, efficient and honest food inspection service. It has day-to-day contact with the food industry, trade and public. The food inspector has many responsibilities, such as: the sampling of food during harvest, processing, storage, transport, or sale to establish compliance with all the stated requirements, the evaluation and implementation of the preventive approaches collected in the Hazard Analysis Critical Control Point System[19] (HACCP) plans, the recognition, collection, and transmission of evidence when breaches of law occur, and so on.

- Laboratory Services (Food Monitoring and Epidemiological data).

  Laboratories appointed to perform both routine and sophisticated/reference food analysis and update food-related databases are an essential key-component of a high-performing food control system. They should have suitable equipment for physical, microbiological and chemical analysis and a full-access to library facilities and data. If more than one laboratory is present on the territory, the tasks can be divided in order to obtain the most efficacious coverage and one could be established for reference analysis. All food analysis laboratories may not be under the control of one agency or ministry, and a number could be under the jurisdiction of the states, provinces, and local authorities. Laboratories must be accredited to ensure reliability, accuracy, and repeatability of the analysis.

- Information, Education, Communication and Training.

  The delivery of information education and advice to stakeholders across the *from farm to table* continuum (end-customers included) has gradually grown in importance for food control systems. Food control agencies should address the specific training needs of their food inspectors and laboratory analysts as a high priority.

[19] Food and Agriculture Organization of the United Nations & World Health Organization. Hazard Analysis Critical Control Point System (HACCP)-Annex of CAC. Retrieved from https://www.fao.org/3/y1579e/y1579e03.htm (last access January 2022).

1.1.3 Food control system-the response of European Union

To ensure the difficult task of a harmonized approach with regard to official controls for verifying compliance with feed and food law, animal health and animal welfare rules, European Parliament and Council released the Regulation (EC) No 882/2004[20]. It gave particular regard to prevent, eliminate, and reduce the risks for humans and animals to acceptable levels, to guarantee fair practices in feed and food trade and to protect consumer interests, including feed and food labeling and other forms of consumer information. It is worth remembering that all the food legislations refer to Regulation (EC) No 178/2002, which set out the general principles and requirements of food law, procedures in matters of food safety, and established the European Food Safety Authority, EFSA.

Since December 2019, Regulation (EC) No 882/2004 has been replaced by Regulation (EU) No 2017/625[21], a legislation globally defined as EU legislation on the agri-food supply-chain (*Union agri-food chain legislation*), which sets the activity of the official control on food safety, extending the scope of the previous regulation to ensure compliance across the European Member States in the Union in the fields of food, feed, animal health and welfare, plant health and plant protection products. The EU, in redesigning the system of official controls to guarantee the application of the legislation on the agri-food supply-chain, leaves the Member States the faculty to apply the specific regulations and to organize the controls in order to verify compliance at all stages of production, processing and distribution. The Regulation (EU) No 2017/625 also applies to sectors not-involved in the previous legislation: plant health, trade and use of protected plant products, organic production, use and labeling of Protected Designations of Origin (PDO), Protected Geographical Indications (PGI), and Traditional Specialty Guaranteed (TSG) products (further details will be provided later in the text). The Regulation establishes that the official controls have to be carried out using appropriate techniques and by well-trained staff to ensure that the Competent Authorities make uniform decisions. The official controls should be regularly carried out with a frequency commensurate with the level of risk involved, but they can also be carried out when there is a suspicion of non-compliance. Official controls by the Competent Health Authorities may also apply to some closely related aspects such as food labeling and consumer information. The Italian Legislative Decree 193/2007[22], for example, identifies the Italian Public Administration bodies assuming the role of Competent Authority in matters of food safety.

---

[20] European Parliament and Council of the European Union. Regulation (EC) No 882/2004. **2004**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32004R0882 (last access January 2022).
[21] European Parliament and Council of the European Union. Regulation (EU) 2017/625. **2017**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32017R0625 (last access January 2022).
[22] Italian Republic. Decreto Legislativo 6 novembre 2007, n. 193 (italian source). **2007**. Retrieved from https://web.camera.it/parlam/leggi/deleghe/07193dl.htm (last access January 2022).

More in details, the main elements of this regulation are:

- the maintenance of a risk-based approach and at the same time the minimization of the burden on operators: the competent authorities, when organizing controls, must consider the previous certifications of the operator's conformity and the reliability of his/her own controls. Furthermore, official controls must be carried out in such a way that the burden of the companies is minimized. It also focuses on frauds: competent authorities must take into account the probability of fraudulent and misleading behaviors when establishing the regularity of the checks, also considering the probability that costumers might be deceived about property, quality, composition or origin of the commodities.

- controls carried out on operators at every step of the *from farm to fork* continuum.

- the requirement of more transparency by enabling the possibility to publish the results of the controls: the transparency of controls is enhanced and the responsibility of the competent authorities towards consumers and operators is increased, thus improving access to information.

- the establishment of EU reference laboratories (EURL).

- the clarification of the "cascade of methods" used in laboratory operations. If there are official EU methods, they must be automatically chosen, otherwise the alternative method must agree with recognized rules or protocols, or it must be developed and endorsed (i.e., validated) by the EURL.

- the definition of guidelines for border controls on imported goods (uniformed directives are applied to border controls on all the products before their access in the EU) and the institution of the Border Control Posts (BCPs).

- the attribution of responsibility to the operator in case of non-compliance with the rules. The list of measures includes, in addition to the pre-existing ones, restrictions or prohibitions of animal movements, its slaughter, if necessary, to defend public and animal health, and the possible closure of the culprit's site.

- the possibility for the Commission to take measures where the control system in one Member State evidently suffers from interruptions that could damage the agri-food chain if the Country has not solved the problem within a certain time.

- the reinforcement of the rules against frauds: to defeat frauds and encourage fair competition.

As previously stated, the new legislation aims to prosecute scammers with severe penalties, thus meaning that the request to devise new analytical methods to detect food fraud will be increasing. A discussion about the fundamental role of the Analytical

Chemistry in this background will be provided in Chapter 2. A brief overview about the issues of food traceability and authenticity is given in Section 1.2.

1.2 Food Traceability

1.2.1 The general framework of Food Traceability

Traceability has been defined by the EU Common Food Law as *"the ability to trace and follow a food, feed, food-producing animal or substance intended to be, or expected to be incorporated into a food or feed, through all stages of production, processing and distribution"*, according to the Article 18 of the EU Regulation No 178/2002. Several traceability definitions have been introduced by different standardization organizations, legislations, and academic literature[23] and there is no common general understanding.

Moreover, food traceability has been classified into numerous typologies in literature. An initial classification [24] is between *forward traceability/tracking* and *backward traceability/trace*. The first one has been defined as the ability to follow the required information (e.g., localization of food) through subsequent stages of the supply-chain top-down towards the final consumer; the second one as the ability to find them (e.g., origin and characteristics of food) through prior stages of the supply-chain bottom-up. Another simple distinction[25] is between *logistics* as the traceability of physical movement of the product and *quality traceability* as the one of safety and quality-related product properties. Traceability has also been defined as *internal*[26], if circumscribed within a link or among food business operators (FBOs), or *external*, when the traceability of product information occurred between links. A more articulated classification[27] distinguishes among: *product*, being the traceability of physical location of food at any stage in the supply-chain; *process*, being the traceability of types, parameters and sequence of activities performed on food throughout the supply-chain; *genetic*, being the traceability of genetic constitution of food; *inputs*, being the traceability of all input raw material properties and suppliers; *disease and pest*, being the traceability of biological hazards of

---

[23] Islam, S.; Cullen, J. M. *Food Control*. **2021**, *123*, 107848.
[24] a) Jansen-Vullers, M.H.; Van Dorp, C.A.; Beulens, A.J.M. *Int. J. Inf. Manage*. **2003**, *23*, 395-413; b) Dupuy, C.; Botta-Genoulaz, V.; Guinet, A. *J. Food Eng*. **2005**, *70*, 333-339; c) Schwägele, F. *Meat Sci*. **2005**, *71*, 164-173.
[25] a) Salampasis, M.; Tektonidis, D.; Kalogianni, E.P. *J. Syst. Inf. Technol*. **2012**, *14*, 302-317; b) Folinas, D., Manikas, I.; Manos, B. *Br. Food J*. **2006**, *108*, 622-633.
[26] a) Borit, M.; Olsen, P. *Mar. Policy*. **2012**, *36*, 96-102; b) Mgonja, J.T.; Luning, P.; Van Der Vorst, J. *J. Food Eng*. **2013**, *118*, 188-197; c) Moe, T. *Trends Food Sci. Technol*. **1998**, *9*, 211-214.
[27] a) Aung, M.M.; Chang, Y.S. *Food Control*. **2014**, *39*, 172-184; b) Karlsen, K.M.; Dreyer, B.; Olsen, P.; Elvevoll, E.O. *Food Control*. **2013**, *32*, 409-417; c) Opara, L.U. *J. Food Agric. Environ*. **2003**, *1(1)*, 101-106.

food; and *measurement traceability*, being the traceability of measured values of parameters.

Traceability has a generic framework, whose fundamental constructs are always valid and independent of the type of product, production, and control system it serves. Figure 1.2, retrieved from literature[23] shows schematically the generic scheme of a Food Traceability System (FTS). The authors states that answering *What? Why? Who? Where?* and *How?* in the proposed framework allows users exploring the definition, drivers, beneficiaries and typologies (keys for understanding the scope of the FTS), and principles (influencing its characteristics) of a FTS.

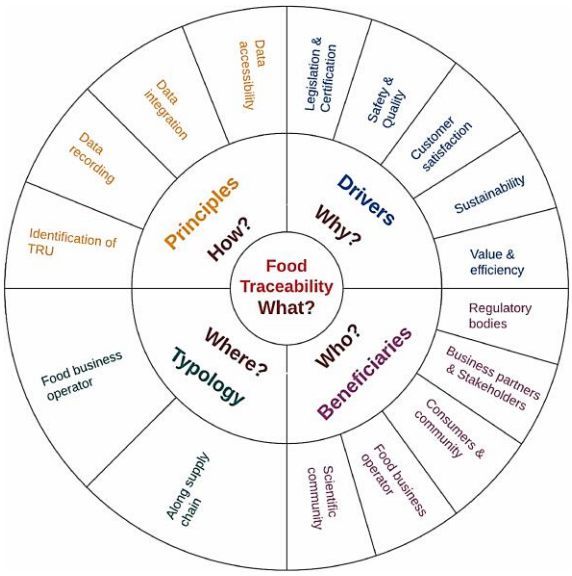


Figure 1.2. Theoretical framework of a FTS (TRU stands for Traceable Resource Unit). Reproduced from Ref. 23.

Traceability principles are the effective operation of FTSs, i.e., the essential characteristics of a system that reflect that system's designed purpose[28]. Many principles have been proposed in literature[29,24a], using often overlapping and conflicting expressions.

---

[28] Alpa, G. *Annu. Surv. Int. Comp. Law*. **1994**, *1*, 1-38.

[29] a) Food Standard Agency (FSA). **2002**. *Traceability in the Food Chain: A preliminary study*; b) Olsen, P.; Borit, M. *Trends Food Sci. Technol*. **2013**, *29*, 142-150.

The recurring ones are:

- identification of Traceable Resource Unit (TRU), i.e., the smallest unit of product to trace, as key operational requirement.

- Recording of data. Food companies decide which data are to be managed by an FTS depending upon why the FTS is required (drivers) and who will use these data (beneficiaries). Under *recording of data* principle, two key decisions are implied, i.e., what traceability data to be recorded and what data recording mediums to be used. The data to be recorded in FTSs are defined by International Organization for Standardization (ISO) and fall under three categories: *shall*, i.e., data which are required for tracing product units; *should*, i.e., data which are required for regulatory or certification needs; and *may*, i.e., additional data required by FBOs for operational purposes. The data recording medium holds the collected data and can vary from a simple paper-based record to a complex computer or cloud-based system. Widely used identification/communication techniques for the benefit of consumers are the smartphone applications, such as the QR code[30] (Figure 1.3) and the Near Field Communication[23] (NFC) technology.

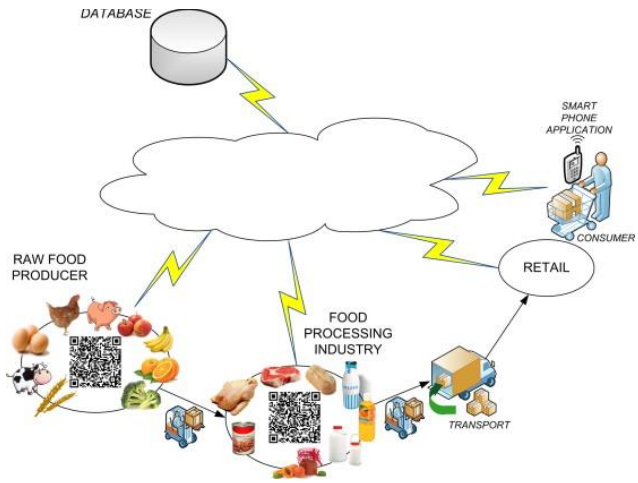


Figure 1.3. Product tracking through the QR code application along all the possible stages in a food chain. Reproduced from Ref. 30.

[30] Tarjan, L.; Šenk, I.; Tegeltija, S.; Stankovski, S.; Ostojic, G. *Comput. Electron. Agric.* **2014**, *109*, 1-11.

Moreover, with the new *European digital strategy*[31] and the digital revolution, emerging technologies such as artificial intelligence, the *Internet of Things* (IoT), sensor technologies, and *blockchain*[32] are being evaluated for their applicability in food production systems and Food Supply-Chain (FSC). These emerging technologies can provide new tools for approaching food safety issues and investigating foodborne incidents, having the potential for improving the speed and precision of the traceability of food products by providing real-time data, which are valuable in outbreak investigations, assessments, and management[33]. The IoT devices combine physical objects, i.e., *things*, with embedded sensors, software and *Internet* connectivity. It could be used as source of information (large amounts of data) as complement to the traditional ones already in place[34], such as surveillance, monitoring, and inspection systems. The large amount of data generated by IoT devices might be integrated with the blockchain technology to increase the reliability of these data[35]. It is a peer-to-peer distributed system in which the stored data become immutable (i.e., cannot be changed) and non-repudiable (i.e., the authenticity cannot be denied). Blockchain could be used to track the records of the journey of a packed food item from the manufactured time to the delivery time of the packet, has been recently proposed as an innovation to address pressing sustainability challenges in global food supply-chain.

- Data integration[36]. The concept of *information integration* encompasses different coordination, linking, merging and sharing of information among different *internal* and *external* FSC partners to achieve a mutually agreed common goal.

- Data accessibility. Different levels of data accessibility have to be provided to internal employees, business partners, legislators and final customers who visualize data in different ways[37]; supply-chain partners indeed are not allowed accessing all traceability data.

---

[31] European Commission. The European Digital Strategy. Retrieved from https://ec.europa.eu/info/publications/EC-Digital-Strategy_en (last access January 2022).

[32] a) Singh, A.; Kumar, V.; Kumar Ravi, A.; Chatterjee, K. *Lect. Notes Electr. Eng*. **2021**, *668*, 745-755; b) Friedman, N.; Ormiston, J. *Technol. Forecast. Soc. Change*. **2022**, *175*, 121403; c) Huan, M.; Ding, Z.; Li, S.; Zhang, C. *Lect. Notes Data Eng. Commun. Technol*. **2021**, *80*, 620-627.

[33] Sarno, E.; Pezzutto, D.; Rossi, M.; Liebana, E.; Rizzi, V. *J. Food Prot*. **2021**, *84*, 2059-2070.

[34] Bouzembrak, Y.; Klüche, M.; Gavai, A.; Marvin, H.J.P. *Trends Food Sci. Technol*. **2019**, *94*, 54-64.

[35] Reyna, A.; Martín, C.; Chen, J.; Soler, E.; Díaz, M. *Futur. Gener. Comput. Syst*. **2018**, *88*, 173-190.

[36] Nakandala, D.; Samaranayake, P.; Lau, H.; Ramanathan, K. *Bus. Process Manag. J*. **2017**, *23*, 108-129.

[37] Šenk, I.; Ostojić, G.; Tarjan, L.; Stankovski, S.; Lazarević, M. *IFIP Adv. Inf. Commun. Technol*. **2013**, *394*, 155-163.

Drivers[38] are the motivating forces, determining *why* a FTS is required:

- Legislation and certification. Traceability has been incorporated into many food legislations, following the standards developed by the ISO[39]. The incorporation of traceability in food regulations gave food safety responsibility to all business operators and actors (governments, producers, and consumers)[27a] through each step of the entire production process, which is viewed as a link in the supply-chain and contribute to spread the traceability issue.

- Safety and quality. Traceability, along with the adoption of the HACCP (see Section 1.1.2), can assist firms to effectively manage safety hazards mitigating safety and quality concerns for consumers[27a].

- Customer satisfaction. More educated customers need to receive more information about the sustainability and source of food from packaging; FTSs may act along the FSC as effective medium, increasing consumer trust[29a].

- Sustainability[40]. Companies can easily verify sustainability claims and mitigate potential reputational risk related to poor environmental or social practices, thanks to FTS.

- Value and efficiency[41]. Traceability helps retailers to improve operational planning, to optimize shelf-life based distribution planning, to prevent animal theft and to minimize product recall.

Five categories of direct beneficiaries of FTSs have been iteratively identified from the food-related literature:

- Public, international standardization and non-governmental certification bodies. All the above-mentioned bodies, in collaboration with stakeholders, contribute to identify important traceability information along the entire value chain in order to develop a credible and robust chain of standards for food, as witnessed by the legislative panorama described in Section 1.1.

---

[38] Thakur, M.; Hurburgh, C.R. *J. Food Eng*. **2009**, *95*, 617-626.

[39] International Standardization Organization (ISO, Geneva, Switzerland). **2007**. ISO 22005:2007: Traceability in the feed and food chain-General principles and basic requirements for system design and implementation. Retrieved from https://www.iso.org/standard/36297.html (last access January 2022).

[40] United Nations Global Compact & Business for Social Responsibility. **2014**. *A Guide to Traceability. A Practical Approach to Advance Sustainability in Global Supply-Chains*. Retrieved from www.bsr.org (last access January 2022).

[41] a) Bourlakis, M.; Bourlakis, C. *J. Enterp. Inf. Manag*. **2006**, *19(4)*, 389-402; b) Smith, G.C.; Tatum, J.D.; Belk, K.E.; Scanga, J.A.; Grandin, T.; Sofos, J.N. *Meat Sci*. **2005**, *71*, 174-193.

- Business partners and stakeholders (exporters, restaurants, and grocery stores). Traceability information collected in FTS improve the ability of supermarkets in optimize sourcing, discount planning, stock rotation, inventory ordering, sales tracking, and the warning of customers who have purchased suspect products[42].

- Consumers and the community[43]. Traceability helps customers to build trust in the food system. Consumers are often willing to pay more for traceable food products and FTS provides detailed information on the environmental and social impacts of products with the aim of helping them making more informed buying decisions.

- Food business operator. Producers often try to differentiate their products from others by providing attributable information (supplied by FTS) on the food product's packaging. Certain credence information, such as Country-of-origin, organic, free-range, or earth-friendly, are particularly valued by customers, providing competitive advantage, increased sales, and higher brand value[44].

- Scientific community. The maintenance of a good food traceability system requires the development of new food labeling and tracking technologies[45]. Furthermore, traceability data are used in numerous research studies to understand traceability performance[26], user experience[46], and consumer expectations[47]. Traceability is nowadays a widely used concept that has proven to be an interdisciplinary complex field developed in different directions both in the food industry and in academic research, spanning different scientific fields like economics and marketing, engineering, social, and natural sciences[27]. The strong interest of the scientific community in food safety and traceability may be easily observed through bibliometric analysis.

A straightforward analysis has been conducted by Sinha and collaborators[48]. They found 805 publications since 1992 to the present day on food traceability topic through Scopus database. A plot of the number of year-to-year and cumulative publications is reported in Figure 1.4. Since 2003 a rising trend in the number of publications can be observed, i.e., academics all over the world have been paying close attention to the subject of traceability in the context of food safety.

---

[42] Golan, E.H.; Krissoff, B.; Kuchler, F.; Calvin, L.; Nelson, K.E.; Price, G.K. Economic Research Service, US Department of Agriculture, Agricultural Economic Report No (AER-830). **2004**. *Traceability in the U.S. Food Supply: Economic Theory and Industry Studies*.

[43] Wu, W.; Zhang, A.; van Klinken, R.D.; Schrobback, P.; Muller, J.M. *Foods*, **2021**, *10(10)*, 2490.

[44] Jedermann, R.; Nicometo, M.; Uysal, I.; Lang, W. *Philos. Trans. R. Soc. A*. **2014**, *372*, 20130302.

[45] Bosona, T.; Gebresenbet, G. *Food Control*. **2013**, *33*, 32-48.

[46] Hernández-Jover, M.; Schembri, N.; Toribio, J.A.; Holyoake, P. *Aust. Vet. J*. **2009**, *87*, 387-396.

[47] Zhang, C.; Bai, J.; Wahl, T.I. *Food Control*. **2012**, *27*, 21-28.

[48] Sinha, A.; Priyadarshi, P.; Bhushan, M.; Debbarma, D. *Artif. Intell. Agric*. **2021**, *5*, 252-261.

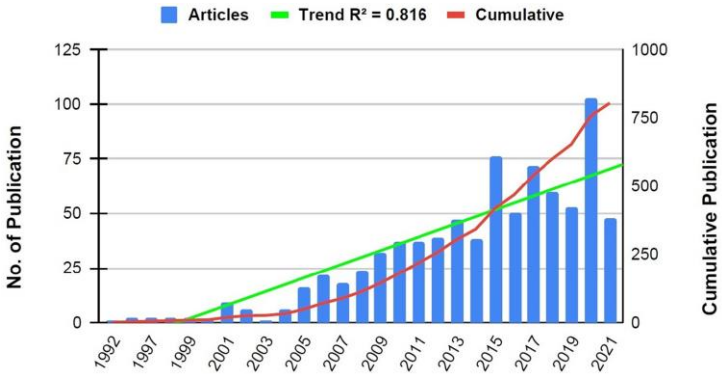


Figure 1.4. Evolution of the number of publications retrieved from Scopus database considering the years 1992-2021. The research papers and conference papers were only chosen. Food-related keywords are used for searching. For further details the reader can refer to Ref. 48, from which the figure is reproduced.

The 805 articles retrieved came from 85 different nations. China had the highest number of publications (199 articles, 18.49% of all papers), followed by the United States of America (USA) with 185 articles (17.19%), and Italy with 77 articles (7.16%). Figure 1.5 gives a graphical representation of the Total Link Strength (TLS) investigation.

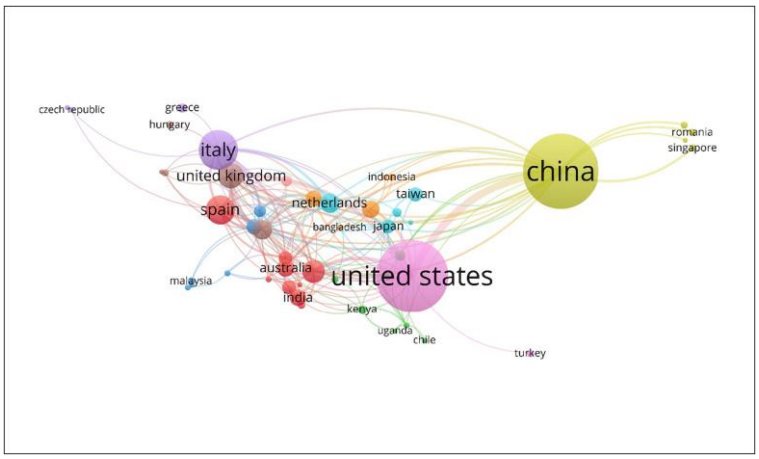


Figure 1.5. Network of Country (considering the ones with at least three documents published) collaborations based on the use of traceability in food safety research. The same information given in Figure 1.3 caption has to be considered. Reproduced from Ref.48.

TLS is a measure of a Country's collaborative research with another. TLS study revealed that the United States, with a TLS of 98, was by far the most superior Country in terms of collaborative research, followed by China (TLS of 85). Italy has a TLS value of 41.

1.2.2 Does Traceability lead to Food Authentication?

Food traceability and food authentication constitute inseparable parts of food safety and quality[49]. FTS is always related to food authentication because it is based on the creditability of originality, i.e., the truthfulness of what is on the label and the presence of contaminants or hazardous substances. The authentication of food is indeed the process by which it is possible to verify if the product complies with the statements on the label and if it is consistent with what is established by regulations; authentication allows recognizing, tackle and prevent adulteration and counterfeiting. The food fraud is a hot topic, as attested by the several reports on adulteration, counterfeiting, and mislabeling published in Europe during the period 2016-2018[50], with a cost for food industry estimated at 8-12 billion € per year. Moreover, the global trade, the new distribution channels and the economic pressure, contribute to food fraud. Many different definitions of Food Fraud have been introduced; the key characteristics in any definition are: the violation of EU food law, the intention, the economic gain, and the customer deception.

---

[49] Dimitrakopoulou, M.E.; Vantarakis, A. *Food Rev. Int*. **2021**, 1-23.

[50] European Commission. *Knowledge Center for Food Fraud and Quality*. Retrieved from https://knowledge4policy.ec.europa.eu/food-fraud-quality/topic/food-fraud_en#Types (last access January 2022).

Figure 1.6. Type of Food Frauds. Reproduced from https://knowledge4policy.ec.europa.eu/food-fraud-quality/topic/food-fraud_en#Types (last access January 2022).

Types of food fraud (schematically reported in Figure 1.6), that can appear alone or in a combination, include (but not limited to):

- *substitution*, i.e., the replacing an ingredient, or part of the product, of high value with another ingredient, or part of the product of lower value;
- *dilution*, i.e., mixing a liquid ingredient of high value with a liquid of lower value (case of study reported in Chapter 10);
- *concealment*, i.e., hiding the low quality of food ingredients or product;
- *mislabeling*, i.e., placing false claim on packaging for economic gain;
- *counterfeiting*, i.e., copying the brand name, packaging concept, recipe, processing method, etc. of food products for economic gain. It can be linked to the general concept of *falsification*, referred to the distribution of a product apparently similar to the high-value one, or false declaration, for example about the composition, processing, qualities, or origin of a product.
- *unapproved enhancement*, i.e., adding unknown and undeclared materials to food products to enhance the quality attributes;
- *grey market production/theft/diversion*, i.e., sale of excess unreported product.

The term *adulteration* is used to involve all these fraudulent practices, together with the food *alteration*, i.e., the natural degeneration of the organoleptic characteristics (rancidity, sourness) and chemical composition of the product as a result of improper processing and/or storage methods. *Substitution*, *dilution* and *unapproved enhancement* can be grouped in the general concept of *sophistication*, i.e., the intentional alteration of the chemical or physical characteristics of a product due to the addition of foreign and/or lower quality substances. It is generally driven by profit and carried out to improve the appearance of the product, shelf-life or to cover up defects.

Food authentication is main concern for official organizations including the scientific community, law enforcement, food producers, importers, exporters, and consumers. This field is in the phase of exponential growth and is attracting a high level of attention from authorities and media around the world. A straightforward analysis[49] has been conducted by the Department of Public Health of the Medical School of the University of Patras (Patras, Greece).

Figure 1.7. Number of publications on food traceability and/or authentication methods according to the food type and origin country. Reproduced from Ref.49.

They carried out a comprehensive literature review on EU-based studies published in the period 2010-2018 using three online databases (Science Direct and Scopus from Elsevier, and PubMed from NCBI) and "Food" AND "traceability" AND "Food" AND "authentication" AND "geographical origin" as search terms. They selected 224 papers, grouped by food type and Country-origin in Figure 1.7.

Seafood, dairy, wines, beverages, plants, honey, meat, and oils were the most common food types used in the included studies in order to confirm the geographical origin or identity the product. The highest percentage, 29.4% (66 publications) presented techniques for identification of geographical or botanical origin of plants, 48 articles (21.4%) report methods for oils' authentication and traceability, 21 published articles (9.3%) were about seafood authentication and traceability, honey adulteration has been recorded in 20 publications due to commercial and health aspects, eighteen reports (8.9%) targeted wine issues, whereas the origin and identity of dairy products is instead studied in 17 publications. The other food types contribute to a lesser extent.

Among the Countries, Italy emerged in the field of food traceability and authentication, being the first Country in publications (37.5%), potentially due to the high number of PDO, PGI and TSG Italian food specialties. Spain and Greece have also contributed with 45 and 20 publications, respectively; while France, England, Germany, and Romania have approximately 10 publications each.

1.2.3 Quality indicators-the international and national panorama

In the scenario illustrated in Section 1.2.1, standardization organizations for controlling the origin of foodstuff and the production process have appeared all over the world.

In 1986 the Organization *Slow Food* has been founded in Italy to promote local food and local small businesses, traditional cuisine, and sustainable foods. Nowadays, it is worldwide spread, and it was the first established part of a broader slow movement: the Slow Food International. Slow Food Italy, in collaboration with Slow Food International, inaugurated in Florence in 2003 the institution of the Slow Food Foundation for Biodiversity[51], to coordinate and promote Slow Food's projects (Presidia, Ark of Taste, Gardens in Africa, Slow Food Chefs' Alliance and Earth Markets, to name a few) to protect food biodiversity across the world. The Foundation, active in over 100 Countries, involves thousands of small-scale producers in its projects, providing technical assistance, training, producer exchanges and communication (guidelines, production protocols, manuals, etc.). The projects of the foundation are tools to provide food sovereignty, to promote a model of agriculture that is based on local biodiversity,

[51] Slow Food Foundation. Slow Food Foundation for Biodiversity. Retrieved from https://www.fondazioneslowfood.com/en (last access January 2022).

respecting the land and the local know-how. Among the projects, the most spread are the Slow Food Presidia that protect the traditional products (together with the Ark of Taste one), the traditional processing methods, and the unique landscape and ecosystems at risk of extinction respecting environmental and social sustainability. Nowadays, more than 600 Presidia, involving 79 Countries, have been set up, and more than 300 of them have been conferred to Italian products. Slow Food Italy registered the *Presidio Slow Food* brand creating an own graphic logo, depicted in Figure 1.8.

Figure 1.8. *Slow Food Presidium* brand instituted by Slow Food Italy.

The marketing of Presidia products has always been carried out directly by producers themselves or producer consortiums. Producers that partake in the project accept both the self-monitoring by their own association and checks by Slow Food Italy whose objective is to ensure the production protocol and general guidelines of the project are respected.

In 2012, Regulation (EU) No 1151/2012[52] of the European Parliament and of the EU Council, on quality schemes for agricultural products and foodstuffs was published. EU quality policy aims to protect the names of specific products to promote their unique characteristics, linked to their geographical origin as well as traditional know-how. Product names can be granted of a *geographical indication* (GI) if they have a specific link to the place where they are made. Products that are under consideration or have been granted GI recognition are listed in quality products registers, such as eAmbrosia[53]. Geographical indications establish intellectual property rights for specific products, whose quality is specifically linked to the area of production. Names of products registered as GIs are legally protected against imitation and misuse within the EU and in

[52] European Parliament and Council of the European Union. Regulation (EU) No 1151/2012. **2012**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=CELEX:32012R1151 (last access January 2022).

[53] European Commission. eAmbrosia-the EU geographical indications register. Retrieved from https://ec.europa.eu/info/food-farming-fisheries/food-safety-and-quality/certification/quality-labels/geographical-indications-register/ (last access January 2022).

non-EU Countries (where a specific protection agreement has been signed). They comprehend:

- PDO–*Protected Designation of Origin* (food and wine);
- PGI–*Protected Geographical Indication* (food and wine);
- GI–*Geographical Indication* (specific for spirit drinks and aromatized wines).

The differences between PDO and PGI are linked primarily to how much of the product's raw materials must come from the area, or how much of the production process has to take place within the specific region. Product names registered as PDO are those that have the strongest links to the place in which they are made: every part of the production, processing and preparation process must take place in the specific region. PGI instead emphasizes the relationship between the specific geographic region and the name of the product, where a particular quality, reputation or other characteristic is essentially attributable to its geographical origin: for most products, at least one of the stages of production, processing or preparation takes place in the region.

Another EU quality label is *Traditional Specialty Guaranteed* (TSG). TSG highlights the traditional aspects, such as the way the product is made or its composition, without being linked to a specific geographical area. The name of a product being registered as a TSG protects it against falsification and misuse.

This Regulation also includes two other optional quality terms, namely Mountain Products and Product of Island Farming.

Some interesting data about registered products comes from the *Study on economic value of EU quality schemes*[54] ordered by the European Commission in 2017. The total sales volume of agricultural products and foodstuffs covered by GIs and TSGs in 2017 was about 10 million tonnes and the total sale value was 27.34 billion €. Excluding TSGs, the sale value of agri-food products under GIs reached 24.95 billion €. PGI products accounted for 54% of the EU sale value in 2017 while PDO and TSG products represented 38% and 9%, respectively. The sector of agricultural products and foodstuffs covered by GIs and TSGs has increased by 39% since 2010. PGI products are the most dynamic and their sales have grown by 60% between 2010 and 2017. TSG products have grown by 50% since 2011 and PDO sales value has increased to a lesser extent by 15% since 2010. In terms of growth contribution between 2010 and 2017, PGI products contributed to 72% of the EU sales values growth followed by PDO and TSG products which represented respectively 18 and 10% of the growth. The four largest categories of agricultural

[54] European Commission-Publications Office of the EU. **2017**. *Study on economic value of EU quality schemes, geographical indications (GIs) and traditional specialties guaranteed (TSGs).*

products and foodstuffs sold under GIs (TSGs excluded) in terms of sales value in 2017 are:

- Cheeses represented more than one third of the total EU-28 (EU constituted by 28 members) sale value for GIs agri-food products. This category has increased by 43% since 2010 and contributed the most to the growth of the entire agri-food sector with 32%.

- Meat products (cooked, salted, smoked, etc. included) accounted for 16% of the total sales value. This category has grown by 24% since 2010 and contributed to 9% of the growth.

- Beers represented 13% of the agri-food sector under GIs in 2017. Beers sales value has grown by 16% but only contributed to 6% of the sector growth since 2010.

- Fresh meats and offal represented 12% of the total EU-28 sale values for agri-food products. It has almost doubled (+97%) since 2010 and is the second largest contributor with 18% of growth.

- Fruits, vegetables and cereals, and fresh fish, mollusks, and crustaceans represent, respectively, 8 and 4% of the total sales value and have a growth rate of 97 and 150% since 2010. These categories contributed to 11 and 8%, respectively, of the growth.

The top three Member States in terms of agri-food GIs sales gathered 65% of the total sales value in 2017 and 47% of the total number of GIs registered were:

- Italy, with 28.1% of the total sales value and 22% of the GIs registered;

- Germany, with 21.3% of the total sales value and 6.8% of the GIs;

- France, with 15.9% of the total sales value and 18.2% of the GIs;

followed by UK, Spain and The Netherlands accounted together for 24.6% of the sales value in 2017 and 20% of the GIs registered.

In Figure 1.9 the Official European logos laid down by the Commission Delegated Regulation (EU) No 664/2014[55] for PDO, PGI, and TSG products are displayed.

[55] European Parliament and Council of the European Union. Commission Delegated Regulation (EU) No 664/2014. **2013**. Retrieved from https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32014R0664 (last access 2022).

Figure 1.9. Official European logos for PDO, PGI and TSG.

About the logos, the EU quality labeling schemes play an important role in the promotion and protection of products with specific characteristics strictly related to the culture and heritage of regions or Countries affirming and valuing the unique socio-cultural and agro-ecological characteristics of a particular place. The potential of logos associated with certain production systems may also be exploited through consumer tasting events, or in farmers' markets for sponsorship purposes. The EU geographical indications system protects the names of products that originate from specific regions and have specific qualities or enjoy a reputation linked to the production territory. Therefore, for all the quality schemes, each EU Country's competent national authority has to take the necessary measures to protect the registered names within its territory. They should also prevent and stop the unlawful production or marketing of products using such a name.

Moreover, the Regulation (EU) No 664/2014 allows Member States introducing own national labeling rules if most consumers attribute significant value to the provision of such information; this is the case of "high value-added" goods, i.e., products with quality brands. In Italy many national marks, in addition to the European ones, are used for the provision of the information that Italian consumer consider explanatory for value and quality, such as *Controlled Designation of Origin* (DOC), *Controlled and Guaranteed Designation of Origin* (DOCG), *Typical Geographical Indication* (IGT) and *Italian Traditional Agri-food Product* (PAT). The latter one was established by the Italian Ministry of Agricultural, Food, Forestry and Tourism Policies and its assignation criteria are mostly focused on the traditions of the different Italian regions.

It is also important to mention that the Product Segregation model[40], outlined in Figure 1.10, is applied in FSC. It implies that certified products have to be physically separated from non-certified ones at each stage along the value chain. This ensures that certified and non-certified materials are not mixed so that the end-product comes from a certified source. At the end, consumers know that 100% of their products consist of certified materials.

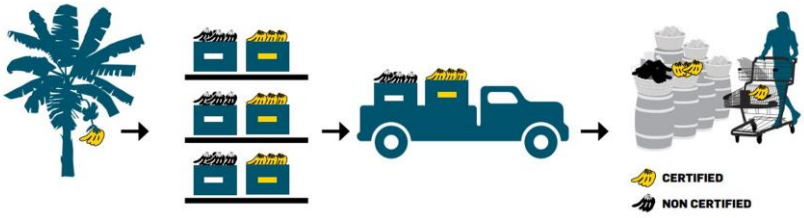


Figure 1.10. Product Segregation model. Retrieved from https://www.bsr.org/ (last access January 2022).

After these considerations, it is clear there is a need to protect high added-value products from adulterations, mislabeling, and, in general, counterfeits. The trade of these products can lead to high profits and, therefore, they can be an object of interest for the scammers.

In this framework, this work intends to present analytical methods to assess the origin, the production, and storage protocol of high-added value foods, focusing on the typical issues of the dairy-products sector.

1.3. Focus on dairy sector

1.3.1 EU regulations on milk and dairy products

The European Union is a substantial producer of milk and milk products, which are integrated in the Regulation (EU) No 1308/2013[56] related to the so-called Common Market Organization (CMO) for the agricultural products. Exceptional ad-hoc measures can be used during periods of severe market disturbance, as the unprecedented situation caused by the COVID-19 crisis, as provided for in Articles 219 to 222 of the CMO. The European Commission indeed approved temporary derogations from EU competition rules, which were available to all farmers across EU Countries. The legal bases for the EU milk and milk products sector are implemented also by two other regulations: the EU Implementing Regulation 511/2012[57] on the milk and milk products sector and EU Regulation 880/2012[58] on transnational cooperation and contractual negotiations of

[56] European Parliament and Council of the European Union. Regulation (EU) No 1308/2013. **2013**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013R1308 (last access January 2022).
[57] European Commission. Commission Implementing Regulation (EU) No 511/2012. **2012**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32012R0511 (last access January 2022).
[58] European Commission. Commission Delegated Regulation (EU) No 880/2012. **2012**. Retrieved from https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32012R0880 (last access 2022).

Producer Organizations (Pos) in the milk and milk products sector. Export licenses and export refunds for milk and milk products must obey to special detailed rules (Commission Regulation (EC) No 1187/2009[59]) as integration of the more general scheme, laying down by Council Regulation (EC) No1234/2007[60]. Similarly, an import regime regulated by Commission Regulation (EC) No 2535/2001[61] is applied for the entrance of the dairy products into the EU.

Several mechanisms (the so-called *Market Intervention*) are used to protect the milk sector during times of increased market disturbance. In particular, market intervention provides a safety net in case of serious market imbalance, in the form of *public intervention* and *aid for private storage*. Public intervention consists of the buying up of a good by public authorities, placing it in public storage for as long as needed, until market conditions allow its releasing back on to the market; in the case of the dairy sector, this is available for butter and skimmed milk powder (SMP). Aid for private storage, instead, supports part of the storage costs while the products are temporarily withdrawn from the market; in the dairy sector, it is available for butter, SMP and PDO or PGI cheeses. Moreover, the *Milk Package*[62] was introduced in 2012, collecting a series of instruments to improve the supply-chain in the dairy sector and to increase its resilience following the end of the *quota* system in 2015. Generally, a production quota is a limit imposed on the quantity produced. Milk quotas were introduced in European Union in 1984 by the Common Agricultural Policy (CAP), following years of significant overproduction of milk and milk products. Two types of quotas were introduced (one defined the maximum amount of milk delivered to dairies and the other the limit for direct sales at farm level) to control the growth of surplus production and budgetary expenditures, to maintain market price support, and to provide price stability for dairy farmers. The milk quotas brought to an end the *milk powder and butter mountains* of the late 1970s and early 1980s, from 1973 to 1983 indeed the production of cow milk on EU farms grew from 92.3 million tonnes to 111.8 million tonnes, a production never again reached.

Milk package stipulates that EU Countries can make written contracts between farmers and processors compulsory and oblige milk purchasers to offer a minimum contract duration to farmers, which may refuse an offer of too short-lived contracts. The contracts should contain specific elements, such as: price, volume, duration, details of payment, collection, and rules for force majeure. Moreover, farmers are encouraged by the EU to

---

[59] European Commission. Commission Regulation (EC) No 1187/2009. **2009**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32009R1187 (last access January 2022).

[60] European Parliament and Council of the European Union. Council 142 Regulation (EC) No1234/2007. **2007**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32007R1234 (last access January 2022).

[61] European Commission. Commission Regulation (EC) No 2535/2001. **2001**. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32001R2535 (last access January 2022).

[62] European Commission. **2017**. *Report from the Commission to the European Parliament and the Council. Development of the dairy market situation and the operation of the 'Milk Package' provisions.*

join together in POs. This is a fundamental right that can partially solve the asymmetry of bargaining power that characterize the dairy sector: there are indeed 11 million farmers in the EU, many of whom work on relatively small family farms which operate independently of each other, and by contrast there is a far higher concentration among both processors and retailers. There are around 3,400 recognized POs in the EU (as of 2017) and they mainly operate in two sectors: fruit and vegetables (52%), and milk and dairy products (9%)[63]. Moreover, specific EU rules for interbranch organizations (IBOs) in the milk sector allow actors in the dairy supply-chain engaging in dialogue and to carry out a number of joint activities, mainly focused on promotion, research, innovation, and quality improvement, with the aim of achieving greater transparency and a better understanding of production and the market.

1.3.2 European Milk Market Observatory

To provide the EU dairy sector with more transparency by means of disseminating market data and short-term analysis in a timely manner, the EU Milk Market Observatory (MMO) was activated.

Farm and dairy herd sizes vary enormously, as do yields. However, as the dairy sector develops throughout the EU, variations in yield and other technical factors have been reduced; less developed dairy producers are rapidly catching up with those who had restructured and modernized first.

Milk production takes place in all EU Countries and represents a significant proportion of the value of EU agricultural output. Total EU milk production is estimated to be more than 155 million tonnes per year (from 158000 to 163250, quantity in 1000 tonnes, in the last five years, from 2017 to 2021), second only to India production (from 176061 to 199000, quantity in 1000 tonnes, within the same time-range). The overall milk production of the main producing countries (India, EU, United States, China, Russia, Brazil, New Zealand, Mexico, Ukraine, Argentina, Canada, Australia, Belarus, Japan, and South Korea, in descending order) ranged from 20 426 to 21 786 (quantity in 1000 tonnes) in the same period. The main producers in the European Union are Germany, France, Poland, the Netherlands, Italy and Ireland, that together account for almost 70% of EU milk production. Italian raw milk production was about 13 million tonnes in 2017[64], with cow milk (12 198,88, quantity in 1000 tonnes) that largely prevails on the production of other niche milk products, such as ewe (480,08, quantity in 1000 tonnes), goat (61,28, quantity in 1000 tonnes) and buffalo (242,99, quantity in 1000 tonnes). Only a small

---

[63] European Commission. **2018**. *Producer and interbranch organizations-Distribution of recognized POs between sectors* (source ISAMM up to 23 July 2018).

[64] Official Journal of the European Union. **2019**. *Communication on the publication of the amounts of raw milk production as referred to in Article 149(5) of Regulation (EU) No 1308/2013 of the European Parliament and of the Council*.

percentage of the milk indeed is from ewes, goats, and buffaloes and it is produced in very specific EU regions. The share of these types of milk grew slightly in the 30 years of application of milk quotas: from 2.3% in 1983 (EU-10) to 3.5% in 2013 (EU-28). In this period the collection of milk from ewes, goats and buffaloes increased by 23% within the EU-10 Member States, reaching close to 3 million tonnes. Greece was the main producing Country of these types of milk with a share of 21.4%, in 2013, followed by Spain (21.1%), France (17.0%), Italy (14.9%) and Romania (12.8%).

Moreover, the EU is a major exporter of dairy products and it is the biggest SMP and cheese exporter in the world. Cheese is, indeed, one of the main dairy products obtained by milk and European Union is the main producer in the world. Its production in the EU in the last five years (2017-2021) ranged from 10 050 to 10 600 (quantity in 1000 tonnes), growing by 3% in the last 12 months[65], while the overall cheese production of the main producing Countries (EU, United States, Russia, Brazil, Argentina, Canada, Mexico, New Zealand, Australia, Belarus, China, Ukraine, Japan, and South Korea, in descending order) ranged from 20426 to 21786 (quantity in 1000 tonnes) in the same period. The extra-EU trade in dairy products has increased in recent years and, according to the European MMO, is expected to continue to grow.

Also in this case, some interesting data about dairy products registered with European quality schemes come from the *Study on economic value of EU quality schemes*[54]. In 2017, the total sale volume of cheeses covered by GIs and TSGs reached more than 1 million tonnes and the value of sales was 9.03 billion € for a total of 241 GIs/TSGs of registered cheese (including 6 TSGs). The sale value of cheeses under GIs (excluding TSGs) reached 8.99 billion € in 2017. Cheeses sale value increased by 43% from 2010. PDO and PGI products equally contributed to the growth of this category of product. Among the 19 Member States producing cheese, Italy, France and the Netherlands are the largest cheese GI producers and accounted for 82% of the total EU cheese sales value in 2017:

- Italy represented 44% of the sale value and 19% of the contribution to growth. The GI cheeses most largely produced in Italy are *Grana Padano PDO*, *Parmigiano Reggiano PDO* and *Mozzarella di Bufala Campana PDO*. In Italy GI cheeses represented 22% of national dairy sector in 2016. This share was relatively stable over the period, though the annual growth rate for GI cheeses was lower (+1.4%/year) than the national growth rate for cheese (+2.8%/year).

---

[65] European Commission. Milk and dairy products-Information on production. Retrieved from https://ec.europa.eu/info/food-farming-fisheries/farming/facts-and-figures/markets/production/production-sector/animal-products/milk-and-dairy-products_en (source: Milk Market Observatory, last access January 2022).

- France accounted for 23% of the sales value and 13% of the contribution to the growth. Five GI products *Comté PDO*, *Roquefort PDO*, *Reblochon PDO*, *Cantal PDO* and *Saint-Nectaire PDO* represented more than half of the volume sales of the French GI cheeses.

- The Netherlands represents 17% of sales value and more than half of the growth observed. The Dutch cheese sector has thrived thanks to the registration of *Gouda Holland PGI* in 2010 and *Hollandse geitenkaas PGI* in 2015.

The GI cheese sector is mainly dominated by cow cheese, as mirror of what happens for raw milk. In 2017:

- Cow cheese was the main subcategory and accounted for three quarters of the EU sales value and growth contribution. The GIs that contributed the most were *Gouda Holland PGI*, *Noord-Hollandse Gouda PDO* and *Parmigiano Reggiano PDO*.

- Mixed milk cheese was the second subcategory with 9% the sales value but contributed to a lesser extent to growth contributing to 5%. The largest GIs are *Feta PDO* and *Kasseri PDO*.

- With 7% of the sales value, sheep cheese contributed to 4% of the cheese sector growth. *Roquefort PDO* is the largest GI for this category of cheese, followed by *Queso Manchegeo PDO*.

- Buffalo cheese accounted for 4% of the sales value and growth contribution. *Mozzarella di Bufala Campana PDO* is the only buffalo cheese under GI in the EU-28.

- Goat cheese sale value reached 4% of the cheese sales value. It increased quickly and contributed to 10% of the growth of the sector, thanks to the registration of *Hollandse geitenkaas PGI*.

EU countries are allowed, under certain conditions, applying rules to regulate the supply of GI (PDO/PGI) cheeses upon the request of a producer organization, an interbranch organization or a PDO/PGI group. This measure is aimed at ensuring the value added and quality of GI cheeses, which are particularly important for vulnerable rural regions. This instrument, known as Regulation of supply of PDO/PGI cheeses is currently applied to: *Asiago PDO*, *Grana Padano PDO*, *Parmigiano Reggiano PDO*, and *Pecorino Romano PDO* in Italy, and *Beaufort PDO*, *Comté PDO*, *Gruyère PGI*, *Reblochon de Savoie PDO*, *Morbier PDO*, *Abondance PDO*, *Emmental de Savoie PGI*, and *Tomme de Savoie PGI* in France.

1.4 Aim and Outline of the work

Food safety and quality authorities are asking for an extended and updated list of analytical techniques for confirmation of food authentication and to support law enforcement. Therefore, there is a growing need for reliable analytical methods that can give a decisive answer about the authenticity of foodstuffs. Food authentication methods must be scientifically proven, accurate and reliable so that food products can be protected and clearly distinguished from any illegal substitutes.

Under this perspective, in the research which constitute the basis of the present PhD thesis, new analytical methods to handle different food traceability and authentication cases in dairy sector will be addressed. More in details, the reader will be led through two Sections dealing with different food-related issues.

Section 1 has to be considered as a single study concerning the multi-platform characterization of three typical Pecorino cheeses produced in Central Italy and certified with European (PDO) and International (Slow Food Presidia) quality schemes. Due to the main characteristics of these dairy products (discussed in the Introduction of Section 1), the study should be considered as an origin and cheesemaking process discrimination issue. Each Chapter (4-7) describes a single-instrument approach, whereas a data fusion approach is presented in Chapter 8, where overall conclusions are drawn.

An abstract of each Chapter of Section 1 is given below.

Section 1

- **Chapter 4**

  HS-SPME/GC-MS volatile fraction determination and Chemometrics for the discrimination of typical Italian Pecorino cheeses[66]

  The volatiles of three typical Italian Pecorino cheeses were analyzed by headspace solid-phase microextraction (HS-SPME) coupled with gas-chromatography (GC) with mass spectrometry detection (MS) evaluating its potentiality to discriminate Protected Designation of Origin (PDO) Pecorino Romano (PR), PDO Pecorino Sardo (PS) and Pecorino di Farindola (PF). A Design of Experiments (DOE) was employed to optimize the extraction conditions in terms of sample temperature and exposure time in the HS-SPME/GC–MS procedure. Linear Discriminant Analysis (LDA) and Partial Least Squares-Discriminant Analysis (PLS-DA) were applied to the volatile composition data to classify samples. Both methods showed high calibration accuracy (100% of non-error rate in cross-validation) and good predictions (87.5% of

[66] Di Donato, F.; Biancolillo, A.; Mazzulli, D.; Rossi, L.; D'Archivio, A.A. *Microchem. J.* **2021**. *165*, 106133.

correct classification, on the external validation set) providing a powerful tool for discriminating Pecorino cheeses, despite their production occurs in different production cycles, in nearby geographical areas and, in the case of PS/PR origins, the raw materials come from common sources.

- **Chapter 5**

Multi-elemental composition data handled by Chemometrics for the discrimination of high-value Italian Pecorino cheeses[67]

The multi-elemental composition of PDO Pecorino Romano, PDO Pecorino Sardo and Pecorino di Farindola, was determined by Inductively Coupled Plasma Optical Emission Spectrometry (ICP-OES). The ICP-OES method here developed allowed the accurate and precise determination of eight major elements (Ba, Ca, Fe, K, Mg, Na, P, and Zn). The ICP-OES data acquired from 17 PR, 20 PS, and 16 PF samples were processed by unsupervised (Principal Component Analysis, PCA) and supervised (PLS-DA) multivariate methods. PCA revealed a relatively high variability of the multi-elemental composition within the samples of a given variety, and a fairly good separation of the Pecorino cheeses according to the geographical origin. Concerning the supervised classification, PLS-DA has allowed obtaining excellent results, both in calibration (in cross-validation) and in validation (on the external test set). In fact, the model led to a cross-validated total accuracy of 93.3% and a predictive accuracy of 91.3%, corresponding to 2 (out of 23) misclassified test samples, indicating the adequacy of the model in discriminating Pecorino cheese in accordance with its origin.

- **Chapter 6**

Fatty acid composition of the triglyceride and free fatty acid fractions in typical Italian Pecorino cheeses

The fatty acids profiles of Pecorino di Farindola, PDO Pecorino Romano and PDO Pecorino Sardo were analyzed by means of GC coupled with flame ionization detection (FID). An extraction-derivatization procedure was *ad-hoc* developed to make possible the analysis of the fatty acids of the triacylglyceride (TAG) and free fatty acid (FFA) fractions of the three high-valued dairy products. To achieve the goal of classifying the samples according to their geographical origin and cheesemaking protocol by exploiting both the information, different strategies of multi-block (MB) data analysis were employed, and the results compared with the ones obtained from PLS-DA analysis performed on one block at the time and taken as benchmark. Eventually, the investigation revealed an increasing of the classification performances when a data fusion approach was applied. Nevertheless,

[67] Di Donato, F.; Foschi, M.; Vlad, N.; Biancolillo, A.; Rossi, L.; D'Archivio, A.A. *Molecules*. **2021**, *26*, 6875.

the GC-fingerprinting results to be a weaker marker of origin and process of production respect to the multi-elemental composition and the volatile profile.

- **Chapter 7**

Application of SPORT algorithm on ATR-FTIR data: a rapid and green tool for the characterization and discrimination of three typical Italian Pecorino cheeses[68]

PDO Pecorino Romano, PDO Pecorino Sardo, and Pecorino di Farindola were analyzed by Attenuated Total Reflectance-Fourier Transformed Infrared (ATR-FTIR) Spectroscopy and classified according to their geographical origin and/or different cheesemaking process. The ATR-FTIR spectra were processed by PLS-DA to be used as benchmark method. Moreover, the multi-block strategy Sequential Pre-processing through ORThogonalization (SPORT), based on Sequential and Orthogonalized Partial Least Squares (SO-PLS), was applied in order to test whether association of spectra pre-processings could enhance the prediction rates. Eventually, the best results were achieved by the multi-block approach, which allows obtaining an accuracy of 98.3% (corresponding to two misclassified samples over 120) in external validation.

- **Chapter 8**

Sequential Data Fusion technique for the characterization and discrimination of typical Pecorino cheeses produced in Central Italy

Final conclusions concerning the classification problem of PDO Pecorino Romano, PDO Pecorino Sardo and Pecorino di Farindola were drawn. The data sets collected in each of the previous chapters were merged to use their joint information in a multi-block framework. SO-PLS-Linear Discriminant Analysis (SO-PLS-LDA) was applied in order to test whether association of different instrumental-blocks could enhance the prediction rates, as well as give an overall characterization of the dairy commodities under investigation. Different merging schemes were applied to give also a straightforward analysis of advantages and drawbacks of *targeted* or *untargeted* analysis application.

The second Section describes the proposal of two different fast, green, and non-destructive approaches implementable *inline* in the FSC to monitor the evolution of different food characteristics during the storage (Chapter 9) and to authenticate niche products (Chapter 10), even subjected to usual technological practices of preservation.

---

[68] Di Donato, F.; Biancolillo, A.; Foschi, M.; D'Archivio, A.A. *submitted to J. Food Compost. Anal.*

An abstract of each Chapter of Section 2 is given below:

Section 2

- **Chapter 9**

Portable 2D-NMR-R technique for the storage monitoring of Pecorino cheeses: a novel approach for the optimization of the acquisition parameter using Chemometrics

The potentiality of low-field 2D ^{1}H-NMR-R as non-destructive method is investigated in the characterization and dating (over three years) of Pecorino cheeses from Central Italy with different ripening times. This work intends to propose a novel approach to set a key-parameter in the acquisition of two-dimensional T_1-T_2 ^{1}H NMR relaxation correlation spectra through unsupervised PCA. Pecorino cheeses were stored under-vacuum and the surface-NMR equipment was used to analyze the sample surface under the packaging. Two-dimensional T_1-T_2 ^{1}H NMR relaxation correlation measurements allowed characterizing the liquid fat and the water phases confined within the cheese matrix. Eventually, ANOVA-Simultaneous Component Analysis (ASCA) was applied to the 2D-data derived from the correlation spectra to investigate potential differences in the bulk-characteristics of Pecorino samples over the years and the ripening time.

- **Chapter 10**

Near Infrared Spectroscopy coupled to Chemometrics for the authentication of donkey milk

Donkey milk is an emerging food commodity which is gaining increasing attention due to its nutritional properties. As a high-added value product, it can be subjected to adulteration with cheaper milks, such as cow milk. The present work addresses the possibility of developing a fast and (relatively) economic method for the authentication of pure donkey milk by means of Near Infrared (NIR) Spectroscopy. On this purpose, 147 samples (67 pure donkey milk and 80 mixtures) were analyzed; additionally, all samples underwent successive freezing-thawing cycles. PLS-DA was used to distinguish pure and adulterated donkey milk samples, before and after freezing-thawing treatment. PLS-DA model on the untreated samples correctly classified all the test samples (100% accuracy); whereas it achieved 81.7% accuracy on freezed-thawed individuals, revealing the effect of adulteration prevails on that of freezing-thawing, even though the latter was proved to be significant by ANOVA-Simultaneous Component Analysis (ASCA).

Abbreviations

B

BCPs, Border Control Posts

BSE, Bovine Spongiform Encephalopathy

C

CAC, Codex Alimentarius Commission

CAP, Common Agricultural Policy

CMO, Common Market Organization

D

DOC, Controlled Designation of Origin;

DOCG, Controlled and Guaranteed Designation of Origin

E

EC, European Commission

EFSA, European Food Safety Authority

EFTA, European Free Trade Association

EU, European Union

EURL, EU Reference Laboratories

F

FAO, Food and Agriculture Organization of the United Nations

FBOs, Food Business operators

FMD, Foot and Mouth Disease

FSC, Food Supply-Chain

FTS, Food Traceability System

G

GFL, General Food Law Regulation

GI, Geographical Indication

H

HACCP, Hazard Analysis Critical Control Point System

I

IBO, InterBranch Organization

IGT, Typical Geographical Indication

IoT, Internet of Things

IPPC, International Plant Protection Convention

J

JECFA, Joint Expert Committee on Food Additives

JEMRA, Joint Expert Meetings on Microbiological Risk Assessment

M

MMO, Milk Market Observatory

N

NFC, Near Field Communication

O

OIE, World Organization for Animal Health

P

PAT, Italian Traditional Agri-food Product

PDO, Protected Designations of Origin

PGI, Protected Geographical Indications

POs, Producer Organizations

R

RASFF, Rapid Alert System for Food and Feed

S

SCCS, Scientific Committee on Consumer Safety

SCHEER, Scientific Committee on Health, Environment and Emerging Risks

SMP, Skimmed Milk Powder

SPS Agreement, Agreement on the Application of Sanitary and Phytosanitary Measures

T

TBT, Technical Barriers to Trade

TLS, Total Link Strength

TSG, Traditional Specialty Guaranteed

W

WHO, World Health Organization

WTO, World Trade Organization

Chapter 2

Analytical Chemistry: a force for food safety and quality

2.1 Introduction

In the framework described in Chapter 1, Analytical Chemistry is strictly involved in the quality control systems, to set up methods as fast, efficient, and cheap as possible, and easily applicable, for example, at a custom level. At the same time, analytical chemistry plays an active role in restoring dignity to traditional food species, due to the potentiality of many available techniques to provide traceability information. The analytical methods represent extremely powerful tools for this purpose, especially when used in combination. The difficulty in food analysis, indeed, is mainly due to the chemical complexity of the food matrix composed of different quantities and classes of compounds, which usually interact and interfere with one another. For this reason, a multi-platform characterization employing different analytical techniques is more advisable, when possible.

As mentioned in Section 1.2.1 and 1.2.2 of Chapter 1, different food traceability and food authentication approaches have been proposed in literature, trying to address the wide range of cases in which food quality control is required. Among them, a distinction in two main categories can be done:

- *targeted/deconvoluted* analysis, i.e., methods for the identification and quantification of a set of target compounds, such as specific known contaminants or metabolites. This approach typically requires a good knowledge of the food matrix under study, with a focus on its production chain and related issues and risks. Target analytical techniques, commonly employed in food authentication, are the DNA-PCR and biosensor methods [69], High [70] and Ultra-High [71] Performance Liquid

[69] a) Montet, D.; Clémentine Kouakou, A.; Hamdouche, Y.; Teyssier, C.; Rychlik, T.; Tatsadjieu, L.N.; Froeder Arcuri, E. In Montet, D. and Ray, R.C. (*Eds.*), *Food Traceability and Authenticity: Analytical techniques* (pp. 137-154). Boca Raton: CRC Press (Food Biology Series), **2017**; b) Bokken, G.C.A.M.; Corbee, R.J.; Van Knapen, F.; Bergwerff, A.A. *FEMS Microbiol. Lett*. **2003**, *222*, 75-82.

[70] D'Archivio, A.A.; Giannitto, A.; Maggi, M.A.; Ruggieri, F. *Food Chem*. **2016**, *212*, 110-116.

[71] D'Archivio, A.A.; Di Donato, F.; Foschi, M.; Maggi, M.A.; Ruggieri, F. *Molecules*. **2018**, *23*, 1851.

Chromatography (HPLC and UHPLC), Gas Chromatography [72] (GC), Mass Spectrometry [73] (MS), and their ifenations, High Resolution-Nuclear Magnetic Resonance[74] (NMR) spectroscopy, and Inductively Coupled Plasma (ICP)-Optical Emission Spectrometry[75] (OES) or ICP-MS[76], to name a few.

This category of techniques is generally characterized by high sensitivity and specificity but presents some critical points. They require optimization steps and sample pre-treatment, increasing the experimental time, cost, and demanding for trained operators.

- *untargeted/convoluted* analysis, i.e., bulk methods for the elaboration of a food fingerprint. Fingerprinting techniques measure the composition of foodstuffs in a non-selective way, producing a large amount of data and variables among which the chemical information useful for the solution of the analytical problem has to be highlighted a-posteriori. Spectroscopic techniques are surely the most commonly exploited when dealing with the delineation of the food fingerprint. Some examples are: Near[77] (NIR) and Mid[78] (MIR) InfraRed spectroscopy, UV-visible spectroscopy[79] (UV-Vis), Raman spectroscopy[80], Hyperspectral Imaging[81] (HSI) and Fluorescence spectroscopy[82]. They show the enormous advantage of being non-destructive and non-invasive (except for UV-Vis in some cases), relatively cheap, rapid, and green; moreover, some of them, NIR spectroscopy above all, can be implementable on-line via probes. As a drawback, most of them are less sensitive and selective with respect to the other ones. Another group of techniques considered as *untargeted* are the ones classified under the term *foodomics*, a discipline introduced in 2009 that integrates untargeted *OMICS* analysis, i.e., metabolomics, proteomics, ionomics and any other type of high-throughput approaches of metabolite profiling generating large amounts of data, in a holistic approach to food and nutrition. Chromatographic methods[83], for example, are widely used for the measurement of metabolomic or/and proteomic profiles (LC-MS, LC-MS/MS, GC-MS, and GC-MS/MS).

---

[72] a) Petrović, M.; Kezić, N.; Bolanča, V. *Food Chem*. **2010**, *122*, 285-291; b) Haglund, P.; Korytár, P.; Danielsson, C.; Diaz, J.; Wiberg, K.; Leonards, P.; Brinkman, U.A.T.; De Boer, J. *Anal. Bioanal. Chem*. **2008**, *390*, 1815-1827.

[73] Di Donato, F.; D'Archivio, A.A.; Maggi, M.A.; Rossi, L. *Food Anal. Methods*. **2021**, *14*, 784-796.

[74] Chatzimitakos, T.; Exarchou, V.; Ordoudi, S.A.; Fiamegos, Y.; Stalikas, C. *Food Chem*. **2016**, *202*, 445-450.

[75] Di Donato, F.; Squeo, F.; Biancolillo, A.; Rossi, L.; D'Archivio, A.A. *Food Control*. **2022**, *131*, 108451.

[76] D'Archivio, A.A.; Giannitto, A.; Incani, A.; Nisi, S. *Food Chem*. **2014**, *157*, 485-489.

[77] Firmani, P.; De Luca, S.; Bucci, R.; Marini, F.; Biancolillo, A. *Food Control*. **2019**, *100*, 292-299.

[78] Di Donato, F.; Di Cecco, V.; Torricelli, R.; D'Archivio, A.A.; Di Santo, M.; Albertini, E.; Veronesi, F.; Garramone, R.; Aversano, R.; Marcantonio, G.; Di Martino, L. *Appl. Sci*. **2020**, *10*, 1630.

[79] Biancolillo, A.; Foschi, M.; D'Archivio, A.A.; Zanardi, E.; Husáková, L.; Hernandez-Hierro, J.M. *Molecules*. **2020**, *25(10)*, 2332.

[80] Firmani, P.; Vitale, R.; Ruckebusch, C.; Marini, F. *Anal. Chim. Acta*. **2020**, *1125*, 308-314.

[81] Caballero, D.; Calvini, R.; Amigo, J. M. *Data Handl. Sci. Technol*. **2020**, *32*, 453-473.

[82] Lenhardt, L.; Bro, R.; Zeković, I.; Dramićanin, T.; Dramićanin, M.D. *Food Chem*. **2015**, *175*, 284-291.

[83] Richard-Forget, F.; Atanasova, V.; Chéreau, S. *Food Control*. **2021**, 121, 107610.

It is important to point out that some of the mentioned techniques (e.g., chromatography, NMR, and elemental analysis) can be used for both *targeted* and *untargeted* analysis.

Regardless of the selected approach, the application of multivariate advanced statistical and mathematical methods is always required. Chemometrics (collecting Data-mining and Machine Learning) is indeed co-evolving fast together with food analysis methods[84]. It is widely known that it has been continuously increasing in food science, especially for monitoring the unit operations, the process, and the quality of food products[85]. A straightforward analysis of the reasons and implications of use of Chemometrics will be given in Chapter 3.

In Section 1, as already stated, the *targeted* analysis of Pecorino cheeses was carried out and compared with its *untargeted* bulk characterization to evaluate the different contribution and potentiality of these two different approaches in a scenario of label-recognition.

In Section 2, two different fast and non-destructive analytical responses were coupled with the same chemometric approach having as final goal the study of the evolution of different food characteristics during their storage.

In this Chapter, an overview of the techniques applied in each of the subsequently reported cases of study is given. The reason why they have been chosen in this thesis will be clarified for each of the investigated cases in the related chapters.

---

[84] Cocchi, M. In Meyers, R.A. (Ed.) *Encyclopedia of Analytical Chemistry* (pp. 1-29). Wiley Online Library, **2017**.
[85] Reid, L.M.; O'Donnell, C.P.; Downey, G. *Trends Food Sci. Technol*. **2006**, *17*, 344–353.

In Table 2.1 the analytical tools are listed, together with an indication of the type of chemometric method to which they have been coupled.

Table 2.1. Outline of the Section 1 and 2.

	Section 1	
Chapter 4	HS-SPME/GC-MS	Experimental design and Classification
Chapter 5	ICP-OES	Classification
Chapter 6	GC-FID	Classification
Chapter 7	ATR-FTIR	Classification
	Section 2	
Chapter 9	2D ^{1}H NMR-R	Exploratory Analysis under Experimental design
Chapter 10	NIR	Exploratory Analysis under Experimental design

HS-SPME/GC-MS: HeadSpace-Solid Phase MicroExtraction/GC-MS
GC-FID: Gas Chromatography-Flame Ionization Detector
ATR-FTIR: Attenuated Total Reflectance-Fourier Transformed Infrared Spectroscopy
2D ^{1}H NMR-R: Two-Dimensional Proton Nuclear Magnetic Resonance Relaxometry

In the following paragraphs is not to provide a detailed discussion about the theory and practice of the well-established above-mentioned techniques will not be given but the main advantages and drawbacks, together with a basic overview will be provided.

2.2 Experimental methods

2.2.1 Gas-Chromatography (GC)

GC[86] is a powerful and widely used analytical technique able to separate even complex mixtures, being included in the rich family of the chromatography. According to the *International Union of Pure and Applied Chemistry* (IUPAC), the modern definition[87] of chromatography applies: "*Chromatography is a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (stationary phase) while the other (the mobile phase) moves in a definite direction*".

In simple terms chromatography can be considered as a series of discontinuous equilibrium steps that take place during a separation. In a small segment of the column

[86] Engewald, W.; Dettmer-Wilde, K. In Engewald, W. and Dettmer-Wilde K. (Eds.) *Practical gas chromatography. A Comprehensive Reference*. (pp. 3-20) Berlin, Heidelberg: Springer, **2014**.
[87] International Union of Pure and Applied Chemistry-Analytical Chemistry Division. Commission on Analytical Nomenclature. **1973**. *Recommendations on nomenclature for Chromatography (Rules approved 1973)*.

(plate) an equilibrium is formed between the solute in the mobile phase and in the stationary phase, which is defined by the solute-specific distribution constant *K*. The chromatographic method using a gaseous mobile phase is called gas-chromatography. The use of a gaseous mobile phase to transport analytes through the column inevitably implies that GC is suitable for non-polar or weakly polar (volatile and semi-volatile) compounds, such as chemicals with less than 60 carbon atoms and a molecular mass and boiling point below 600 Da and 500°C, respectively. GC is indeed often used for the analysis of sterols, oils, aroma components and off-flavors, and many contaminants, such as pesticides, industrial pollutants, and certain types of drugs in foods[88]. Nevertheless, some high-boiling compounds, such as long-chain saturated hydrocarbons up to a carbon number of 100, or triglycerides, can be measured by GC due to their thermal stability. Moreover, many applications report the use of chemical derivatization to enlarge the domain of GC. There are several options to make higher-boiling compounds amenable to GC analysis. Polar, less volatile, or thermally fragile substances can be transformed into stable, volatile, and less polar derivatives by chemical reaction, such as silylation, alkylation, or acylation.

The required analyte transfer into the gas phase, often considered as a disadvantage of GC, offers an elegant option to combine sample introduction with an analyte/matrix separation (during sample pre-treatment). Sample preparation is considered as a bottleneck in the overall analytical procedure, being the most time-consuming step; moreover, errors occurring in this primary step of the analytical process can hardly be corrected. Sample preparation prior to GC separation should aim at the following goals: the improvement of the gas chromatographic behavior of the analytes and analyte detectability, the separation of analytes from interfering matrix components and the enrichment of the target analytes. Classical sample preparation methods are *liquid–liquid extraction* (LLE) for liquid samples and *Soxhlet extraction* for solid samples. Both are simple and largely applied, but they typically need large volumes of sample and organic solvents. An alternative to LLE extraction is *solid-phase extraction* (SPE). In water analysis, SPE is nowadays the most widely used sample preparation technique for non-polar compounds. Survey of the recent literature reveals that extraction, enrichment and matrix separation show major trends towards miniaturization, online sample preparation, automation, high sample throughput by lower processing times, higher extraction yields and better reproducibility, as well as organic solvent reduction or exclusion.

In this context a number of solvent-free sample preparation and enrichment techniques have been developed. In 1989, Belardi and Pawliszyn introduced the concept of *solid phase microextraction*[89] (SPME), which became commercially available in 1993. It combines a solvent-free analyte extraction by a coated silica fiber with the subsequent thermal

---

[88] Lehotay, S.J.; Hajšlová, J. *TrAC Trends Anal. Chem*. **2002**, *21*, 686-697.
[89] Belardi, R.P.; Pawliszyn, J.B. *Water Pollut. Res. J. Canada*. **1989**, *24*, 179-191.

desorption/evaporation of the analytes from the fiber in the hot GC injector. Nowadays, SPME and its fashions are an efficient answer to the above-mentioned requests, combining several sample preparation steps such as sampling, enrichment and clean-up. A thin fused silica fiber coated with extraction phase (7–100 μm film thickness and typically 1 cm length) is utilized for the extraction of analytes from the sample. Many different kinds of SPME coatings are available in commerce, covering a wide range of chemical compounds. Fibers can be used in two application modes. One mode is Direct Immersion (DI) of the fiber into the sample (DI-SPME, Figure 2.1(A)); the other one is the analyte extraction from the headspace (HS-SPME, Figure 2.1(B)) above the sample. After the extraction, desorption is performed thermally in the GC injector; the fragile fiber is encompassed in a special syringe holder for its protection during penetration of the vial and GC injector septa.

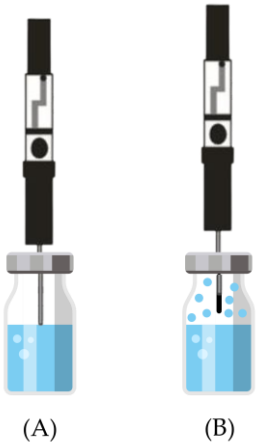


Figure 2.1. Two extraction mode: (A) Direct immersion solid-phase microextraction (DISPME); (B) Headspace solid-phase microextraction (HS-SPME).

SPME is the most prominent and widely used microextraction method with hundreds of publications and applications. Nevertheless, so far only few standardized procedures have been implemented[90].

---

[90] a) International Standardization Organization (ISO, Geneva, Switzerland). **2010**. DIN-EN-ISO-17943: *Water quality-Determination of volatile organic compounds in water-Method using headspace solid-phase micro-extraction (HS-SPME) followed by gas chromatography-mass spectrometry (GC-MS)*. Retrieved from https://www.document-center.com/standards/show/DIN-EN-ISO-17943 (last access January 2022); b) United States Environmental Protection Agency (EPA). **2007**. Validated Test Method 8272: *Parent and Alkyl Polycyclic Aromatics in Sediment*

A gas-chromatograph, schematized in Figure 2.2, consists of six main components.

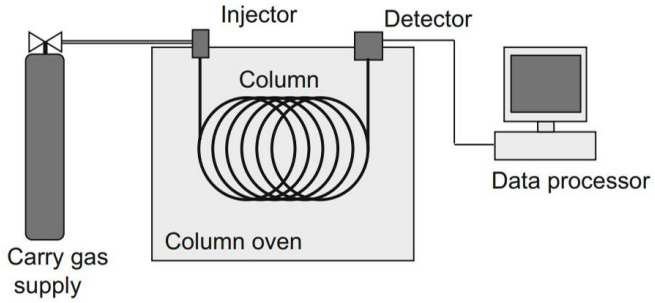


Figure 2.2. Simplified scheme of a typical GC system. Reproduced from Ref.86.

The *carrier gas supply* provides a continuous flow of the mobile phase (carrier gas) through the column. Ultrapure helium, hydrogen, or nitrogen are used as carrier gases. Hydrogen can also be conveniently and safely produced on-site by a hydrogen generator, through the electrolysis of water. The *injector* is a device to introduce gaseous or liquid samples onto the column head. The *column*, mainly capillary ones, is the heart of the GC system, allowing the separation of analytes. The column oven houses the column. The *oven* is an air thermostat to provide a constant (isothermal) or defined (programmed) increased of the column temperature. The *detector* is a device to record the solutes upon leaving the column. The most widely used detectors are the *flame ionization detector* (FID) and the *electron capture one* (ECD), both presented in 1958 during the *International Symposium on Chromatography* (ISC) in Amsterdam. FID has become an extremely powerful device for almost all organic compounds and it is nowadays used as standard equipment in commercial gas-chromatographs. It provides lower LOD, higher robustness, and fast response rates compared to the previously used thermal conductivity detector. The ECD allows the selective and sensitive detection of analytes with a high electron affinity, such as compounds containing halogen atoms, or aromatics, feature that makes it important for the application of GC in environmental analysis. An electric signal is produced by the detector, in most cases amplified, and sent to the data system. The *data processor* is used to register, store, and analyze the data produced.

An important technological progress was the hyphenation of GC to MS. The online coupling of GC-MS was first performed in 1956 with packed columns and later with

---

Pore Water by Solid-Phase Microextraction and Gas Chromatography/Mass Spectrometry in Selected Ion Monitoring Mode. Retrieved from https://www.epa.gov/hw-sw846/validated-test-method-8272-parent-and-alkyl-polycyclic-aromatics-sediment-pore-water-solid (last access January 2022).

capillary ones. The advantages of MS over other GC detectors are the high detection sensitivity, the potential to use it as universal or mass-selective detector, and the option to gain structural information from the mass spectra. Efficiency and robustness of the instruments were constantly improved and mass spectral libraries extended. In addition to quadrupole and ion trap mass analyzers, triple-quadrupole instruments and time-of-flight systems are nowadays used for GC-MS also in routine analysis.

At this stage, gas chromatography presents itself as a widely used and matured technique. Currently, it is estimated that about 400000-500000 GC systems are in use worldwide. Still, the development of GC has not reached its end, but advances are to be expected. Hot topics are high sample throughput, reduction of analysis time, improving of trace analysis, integration of sample preparation and enrichment within the GC instrument, and further automation, in particular in data analysis.

2.2.2 Multi-elemental profile-Inductively Coupled Plasma-Optical Emission Spectrometry (ICP-OES)

Since its commercial introduction in the mid-1970s, ICP-OES[91], also known as inductively coupled plasma-*atomic* emission spectrometry (ICP-AES), has rapidly become widely used and accepted for many applications of metal determination in a wide variety of samples [92]. It may be listed among the analytical techniques based on atomic spectrometry, involving, as the name implies, electromagnetic radiation that is adsorbed by and/or emitted from atoms of a sample. In OES, samples are subjected to high temperatures causing both dissociation into atoms and a significant amount of collisional excitation (and ionization) of the atoms. This thermal excitation source populates a large number of different energy levels for several different elements at the same time. Once the atoms and ions are in their excited states, they can decay to lower ones through thermal or radiative (emission) energy transitions, emitting their characteristic radiation at nearly the same time. This results in the flexibility to choose from several different emission wavelengths for an element and in the ability to measure emission from several elements concurrently. However, the probability for interferences (arising from emission lines too close in wavelengths to be measured separately) can increase with the increasing of the number of emission wavelengths. Among the electrical discharges (a current/potential is applied across an electrode in an inert gas) as thermal sources, the argon-supported ICP is the mainstay of OES. Moreover, because ICP can efficiently

---

[91] Boss, C.B.; Fredeen, K.J. Concepts, Instrumentation and Techniques in Inductively Coupled Plasma Optical Emission Spectrometry. Shelton, CT, US: PerkinElmer Life and Analytical Sciences, **2004**.
[92] Sneddon, J.; Vincent, M.D. *Anal. Lett*. **2008**, *41*, 1291-1303.

generate singly charged ions, it makes an ideal source also to use synergistically with mass spectrometry[93] (ICP-MS).

ICP-OES is nowadays applied in many fields like geology, forensic science, biology, food analysis, and industry, thanks to its ability to analyze liquid, solid and gaseous samples rapidly performing qualitative and quantitative multi-elemental analyses with relatively low limits of detection (LOD). Moreover, due to the wide range of linearity, these technique permits to simultaneously quantify major and trace elements[94].

The sample introduction system is schematized in Figure 2.3.

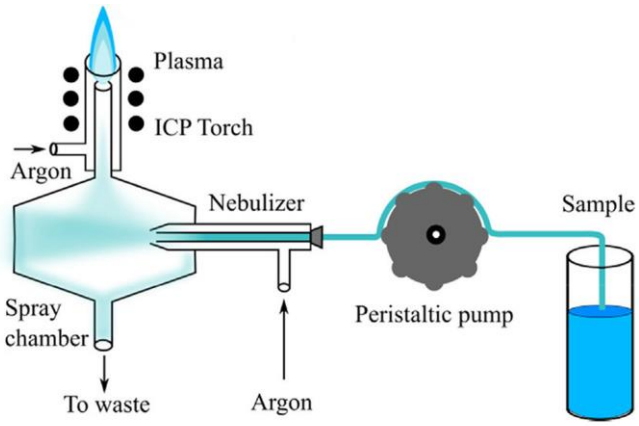


Figure 2.3. A simplified scheme showing the introduction system of ICP-OES. Adapted from Ref.95.

In ICP-OES, the sample is usually transported into the instrument as a stream of liquid sample, maintained by a *peristaltic pump*. Inside the instrument, the liquid is broken up into an aerosol through a process known as *nebulization*. The sample aerosol is then transported to the plasma. Most commercial ICP *nebulizers* are of the pneumatic type, as the one in Figure 2.3, i.e., nebulizers that use high-speed gas (i.e., argon) flows to create an aerosol, "shattering" the liquid into tiny droplets. Because the plasma discharge is not very efficient at dissociating large droplets, it is crucial allowing only the small ones to enter the plasma. The *spray chamber* is designed for this purpose. The *cyclonic spray chamber*, discriminating droplets according to their size, is the most employed for its

[93] Thomas, R. Practical Guide to ICP-MS: A Tutorial for Beginners, Second Edition (2nd ed.). Boca Raton: CRC Press, **2008**.
[94] Wang, T. In Grinberg, N. and Rodriguez, S (Eds.) *Ewing's Analytical Instrumentation Handbook*, Fourth Edition (pp. 57–74). Boca Raton: CRC Press, **2018**.

higher sampling efficiency, sensitivity, and lower detection limits[95]. Indeed, a decrease in the size of aerosol particles permits to minimize the contribution of the solvent, which causes an increase in noise and matrix effects. In addition, the physical properties of the sample matrix (surface tension, volatility, density and viscosity) influence the characteristics of the aerosol. Sample introduction is therefore a crucial step of the analysis. In 1984, indeed, Browner and Boorn[96] described sample introduction as the "Achilles' Heel of Atomic Spectroscopy".

Shortly thereafter, Barnes[97] qualified instead the sample preparation as the limiting factor in atomic spectroscopy. In practice, despite the passing of about 30 years, both these statements are true today. Nevertheless, many developments there have been. The use of microwave technology, for example, has greatly improved sample preparation[98].

Once the nebulizer gas brings the sample from the sample introduction system to the plasma, the emission radiation from the plasma has to be sampled. The radiation is usually collected by focusing optic (e.g., convex lens or concave mirror) that focuses the image of the plasma onto the entrance slit of a spectrometer, i.e., wavelength dispersing device. The discrimination of the emission may be done in several ways; the physical dispersion of the different wavelengths by a *diffraction grating* is by far the most common. Another dispersion device is the prism. In recent years, the advantages coming from the combination of two dispersing systems (a diffraction grating and a prism, or two diffrating gratings) have been shown. The *Echelle grating*, an additional dispersive device that has become an important component in ICP instruments for its high resolving power and flexibility, is usually one of them. The echelle grating separates the polychromatic radiation producing multiple, overlapping spectral orders. The second dispersing device separates or cross disperses the overlapping orders into a two-dimensional pattern called *echellogram*. A typical optical configuration for this echelle type of spectrometer is illustrated in Figure 2.4.

---

[95] Cherevko, S.; Mayrhofer, K.J.J. *Encycl. Interfacial Chem. Surf. Sci. Electrochem*. **2018**, 326-335.
[96] Browner, R.F.; Boorn, A.W. *Anal. Chem*. **1984**, *56*, 786A-798A.
[97] Barnes, R. *Spectroscopy*. **1986**, *1*, 24-32.
[98] Sneddon, J.; Hardaway, C.; Bobbadi, K.K.; Reddy, A.K. *Appl. Spectrosc. Rev*. **2006**, *41*, 1-14.

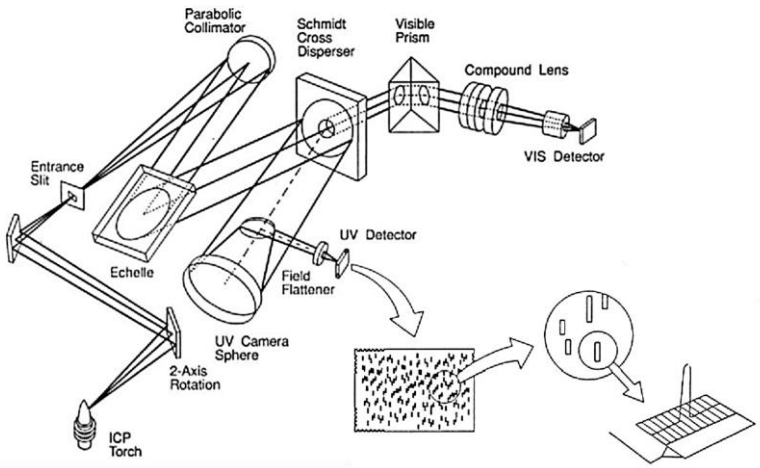


Figure 2.4. Layout of the optical system in a Echelle spectrometer. Adapted from Ref 99.

Echelle spectrometers are generally connected to charge-transfer detectors (CTD). CTDs are silicon-based devices using the light-sensitive properties of solid-state silicon. They are characterized by high sensitivity and resolution and comprise the photodiode array (PDA), the charge-injection device (CID) and the charge-coupled device (CCD). These systems are multichannel devices capable of generating a two-dimensional array of more than 20 thousand individual elements, called *pixels*, operating as detector. The total spectrum is divided into two output regions, UV and visible, which are optimized independently for resolution and throughput[99]. Each region has its own cross-dispersion, focal length, and multichannel solid-state detector. The atomic emission spectrum appears as a series of narrow peaks on a noisy background, which arises from energy losses by radiative decay of free particles, molecular bands, or continuous radiation. The improving of the accuracy and precision of the method is however possible performing correction of the signal, using appropriate internal standards, or acting on the operating parameters for method optimization.

[99] a) Barnard, T.W.; Crockett, M.I.; Ivaldi, J.C.; Lundberg, P.L. *Anal. Chem*. **1993**, *65*, 1225-1230; b) Barnard, T.W.; Crockett, M.I.; Ivaldi, J.C.; Lundberg, P.L.; Yates, D.A.; Levine, P.A.; Sauer, D.J. *Anal. Chem*. **1993**, *65*, 1231-1239.

2.2.3 Fourier Transform-Infrared (FT-IR) Spectroscopy

Spectroscopic techniques exploit the interaction between electromagnetic radiation and matter to characterize the samples. Due to the many possibilities offered by the electromagnetic spectrum in terms of the frequency ranges that can be explored and the various modalities that can be employed to register the signal (e.g., reflection, transmission, transflection), spectroscopic techniques are highly versatile tools for the characterization of foodstuffs.

In analytical chemistry, infrared spectroscopy has always been widely used, but, in recent years, it has been in the bulls-eye and it is considered nowadays a solid alternative to the more traditional analytical methods (i.e., chromatographic techniques), being easy to use, rapid, green (solvent-free), non-destructive and non-invasive, cheap, exploitable for in-situ analysis and also implementable for process-monitoring in industrial applications.

Briefly, infrared spectroscopy[100] deals with the absorption that takes place in the spectral range between 13000 and 10 cm^{-1}. In this domain, it is possible to observe the vibrational modes (hence the name of *vibrational spectroscopy*) such as stretching and bending. The infrared spectral range is sub-divided into three different regions: NIR (13000-4000 cm^{-1}), MIR (4000-200 cm^{-1}), and far infrared (FIR, 200-10 cm^{-1}). Historically, the wider investigated range is the MIR; nevertheless, in the last few years, NIR spectroscopy has been "rediscovered" thanks to the development of data analysis tools.

2.2.3.1 Near InfraRed (NIR) Spectroscopy

The absorption bands in NIR[101] spectra derive from *overtones* and *combinations* of fundamental vibrations modes of organic molecules; due to the highly anharmonic nature of such transitions it mostly involves X-H bonds, where X can be carbon, nitrogen, oxygen, and sulfur. *Overtone* corresponds to the transition from the ground state to higher excited states and occurs at multiples of the fundamental transition vibration frequency. *Combination bands* arise from the simultaneous occurrence of two or more transitions and occur at sums of multiples of the frequencies of each interacting mode. Based on the nature of the transitions involved, which have a lower probability of occurrence than their fundamental counterparts, NIR bands are very weak (10 or 100 times less intense) and typically broad and highly overlapping, with consequent high multicollinearity. Moreover, water, being the main constituent of the majority of foods, strongly absorbs the NIR radiation so commonly prevail on other compounds' peaks. For these reasons, the assignment of peaks to specific chemical constituents is not

---

[100] Lohumi, S.; Lee, S.; Lee, H.; Cho, B.K. *Trends Food Sci. Technol*. **2015**, *46*, 85-98.

[101] Menezes, J.C.; Ferreira, A.P.; Rodrigues, L.O.; Brás, L.P.; Alves, T.P. In Brown, S.D.; Tauler, R. and Walczak, B. (Eds.) *Comprehensive Chemometrics*, Volume 4 (pp. 313-355), Amsterdam, The Netherlands: Elsevier, **2009**.

straightforward. These represent the principal drawbacks of NIR spectroscopy and indeed it is generally inherently less sensitive and selective than chromatography and other destructive analytical tools. Furthermore, the *Lambert and Beer law*, correlating the concentration of a molecule with its absorbance, is not directly applicable to turbid and solid media, such as the food samples. The light entering the sample can be back-scattered (diffuse reflectance) or scattered forward (diffuse transmission); the absorbance can take place directly or following dispersion: reflected radiation can be collected to subsequently obtain absorbance spectra (more precisely, pseudo-absorbance). Consequently, as well as to provide a fingerprint of its chemicals, spectroscopic information of the physical properties of the sample under study is also captured.

Notwithstanding, the coupling with chemometric methods (even more necessary for NIR than with other spectroscopies) has allowed maximizing the extraction of information from NIR signals, and nowadays, it is one of the most widely used techniques, in particular in food quality analysis[102]. This is also due to the fact that these same characteristics (just described as drawbacks) are also responsible for the main advantages of NIR with respect to other spectroscopic techniques: indeed, the low absorptivity allows a higher penetration depth, which in turn results in the possibility of analyzing thicker, strongly absorbing or, even, highly scattering samples, without any need of sample pre-treatment.

In response to the need for speed in analyses and flexibility in adapting to different sample states, NIR spectroscopy instrumentation has evolved accordingly. This translates in a higher flexibility of NIR spectrophotometers that, despite essentially identical with those employed in other regions of the electromagnetic spectrum, can incorporate a variety of devices depending on the characteristics of the sample and the particular analytical conditions and needs[103].

Nearly all modern MIR spectrometers are Fourier transform (FT) based instruments with different interferometer designs (Michelson or polarization interferometers, acousto-optic tunable filters (AOTF)). In contrast, monochromator/detector principles in scanning NIR spectroscopy are variable. A typical NIR spectrometer (Figure 2.5) consists of a radiation source, a wavelength selection device (e.g., a monochromator), a sample holder, a photoelectric detector for the measurement of the intensity of the detected light and conversion into electrical signals, and a computer system for spectral data acquisition and processing[104].

---

[102] a) Biancolillo, A.; Marini, F. In Lopes, J. and Sousa, C. (Eds.) *Vibrational Spectroscopy for Plant Varieties and Cultivars Characterization* (pp. 69-104) *Comprehensive Analytical Chemistry*, Volume 53, Amsterdam, The Netherlands: Elsevier, **2018**; b) Hassoun, A.; Shumilina, E.; Di Donato, F.; Foschi, M.; Simal-Gandara, J.; Biancolillo, A. *Molecules*. **2020**, *25*, 4472.

[103] Blanco, M.; Villarroya, I. *TrAC Trends Anal. Chem*. **2002**, *21*, 240-250.

[104] Siesler, H.W. In Burns, D.A. and Ciurczak, E. (Eds.) *Handbook of Near-Infrared Analysis* (pp. 7-19), Boca Raton: CRC Press, **2008**.

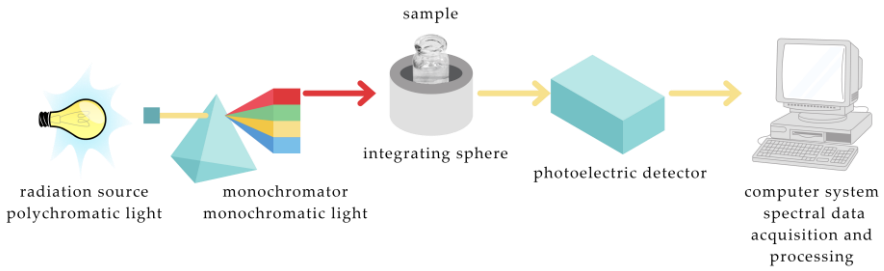


Figure 2.5. Schematic representation of the FT-NIR main components.

Different optical geometries are available for NIR spectroscopy. Predominant sample presentation modes being distinguished are *transmission*, *reflection*, *transflection*, and *interaction*. According to the reflection and transmission optical geometry used, light attenuation by the sample, relative to the reference, is referred to as reflectance (R) and transmittance (T), respectively. Most studies use log 1/T or log 1/R values before performing statistical analysis[105].

2.2.3.2 Mid InfraRed (MIR) Spectroscopy

MIR[106] deals with the fundamental vibrational transitions and results in relatively narrow bands, which can be associated either to the presence of specific functional groups or to the movement of the whole molecular backbone. In this respect, MIR spectroscopy can be advantageously used to characterize, identify and quantify. Indeed, Mid-InfraRed (IR) spectroscopy is an extremely reliable and well recognized fingerprinting method. One of the strengths of IR spectroscopy is its ability as an analytical technique to obtain spectra from a very wide range of solids, liquids and gases. However, in many cases some form of sample preparation is required in order to obtain a good quality spectrum. The technique of Attenuated Total Reflectance[107] (ATR) has in recent years revolutionized solid and liquid sample analysis because it overcomes the most challenging aspects of infrared analyses, namely sample preparation and spectral reproducibility. An attenuated total reflection accessory (Figure 2.6) operates by

---

[105] Herold, B.; Kawano, S.; Sumpf, B.; Tillmann, P.; Walsh, K. In Zude, M. (Ed.) *Optical Monitoring of fresh and processed agricultural crops* (pp. 141–249), Boca Raton: CRC Press, **2009**.
[106] Sun, D.-W. Infrared spectroscopy for food quality analysis and control. Amsterdam, The Netherlands: Elsevier, **2009**.
[107] Ramer, G.; Lendl, B. In Meyers, R.A. (Ed.) *Encyclopedia of Analytical Chemistry* (pp. 1–24). John Wiley & Sons, Ltd, **2013**.

measuring the changes that occur in a totally internally reflected infrared beam when the beam comes into contact with a sample.

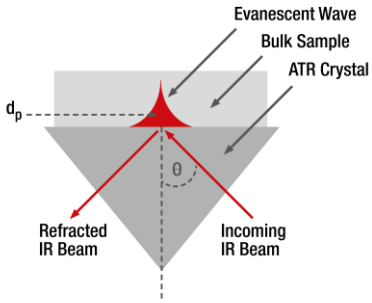


Figure 2.6. Schematic representation of the ATR accessory. Retrieved from https://wiki.anton-paar.com/tw-zh/attenuated-total-reflectance-atr/ (last access January 2022).

An infrared beam is directed onto an optically dense crystal (parallel-sided plate in the horizontal ATR units, HATR) with a high refractive index at a certain angle. This internal reflectance creates an evanescent wave that extends beyond the surface of the crystal into the sample held in contact with the crystal. This evanescent wave protrudes only a few microns (0.5 μ-5 μ) beyond the crystal surface and into the sample. Consequently, there must be good contact between the sample and the crystal surface. In regions of the infrared spectrum where the sample absorbs energy, the evanescent wave will be attenuated or altered. The attenuated energy from each evanescent wave is passed back to the IR beam, which then exits the opposite end of the crystal and is passed to the detector in the IR spectrometer. The system then generates an infrared spectrum. There are a number of crystal materials available for ATR: zinc-selenide (ZnSe) and Germanium are by far the most common used for HATR sampling; diamond is by far the best ATR crystal material, due to its robustness and durability.

2.2.4 Two-Dimensional Proton Nuclear Magnetic Resonance Relaxometry (2D ^{1}H NMR-R)

NMR spectroscopy offers high potential for the analysis of multicomponent systems, such as food products; moreover, the coupling with chemometrics allowed exposing the hidden relationships in complex data of food systems[108]. This growing interest of the application of NMR spectroscopy to solve particular problems in food science has been duly recorded by the increasing number of applications of this technique by food researchers[109] and by a series of biannual conferences devoted to *Applications of magnetic resonance in food science*, that started in 1992 and continue to date. NMR areas of expertise comprise high-resolution liquid- and solid-state NMR spectroscopy, magnetic resonance imaging (MRI), relaxometry, and diffusometry. These methods are different in many aspects, such as instrumentation, sample preparation, and targeted results; nevertheless, they share the same magnetic resonance principle: a collection of nuclei of atoms with magnetic properties is distributed onto various energy levels defined by the orientation of their magnetic moments with respect to an external magnetic field (either homogeneous or inhomogeneous depending on the specific technology). After reaching the so-called *thermal equilibrium*, nuclei are irradiated by a second weak radiofrequency field. The excited nuclei give back their excess energy and return to low energy levels by two relaxation processes. The first one, called spin-lattice relaxation, consists of the interaction with the environment (lattice), i.e., with the other molecules in the system, and is characterized by a time constant T_1, known as the spin-lattice (or longitudinal) relaxation time. The second process, i.e., the exchange of energy with neighboring nuclei (with priority to the ones of the same molecule) at lower energy levels is the spin-spin relaxation; it is described by a time constant T_2, known as the spin-spin (or transversal) relaxation time. Due to the physical origins of these two phenomena arising from the energy transfer from the excited nuclei to the other dynamics in the system for T_1 constant and the redistribution of the energy of the excited nuclei into the molecular degrees of freedom for T_2 constant, T_1 is always higher than T_2, i.e., T_1 relaxation process is the slower one. Relaxation mechanisms involve local magnetic fields produced by the atomic and electronic environment of the nucleus and modulated by molecular motion.

Apart from static properties, food materials show time-dependent phenomena reflected on their structural and dynamic heterogeneities. Some representative examples are: phase transitions, matrix and water distribution, and compositional changes during processing and storage (aging). The study of these time-dependent progressions requires time-domain (TD) NMR experiments (opposed to frequency-domain measurements). They can be carried out at low magnetic fields using low-cost NMR spectrometers, suitable for the industrial sector performing on-line and automated quality assessment

---

[108] Dais, P.; Spyros, A. In Picò, Y. (Ed.) *Chemical Analysis of Food: Techniques and Applications* (pp. 91-115), Amsterdam, The Netherlands: Elsevier, **2012**.
[109] Tang, F.; Vasas, M.; Hatzakis, E.; Spyros, A. *Annu. Reports NMR Spectrosc*. **2019**, *98*, 239-306.

of food products. The parameters for monitoring time-dependent phenomena are precisely the above-mentioned relaxation times and the self-diffusion coefficients.

The development of the first commercial dedicated pulsed TD-NMR instrument started in 1970 with an industrial collaboration between the instrument manufacturer Bruker and a food company (Unilever). In the decade that followed, the range of relaxometric methods widened because the food scientists in academia and industry rapidly recognized the potential of these relatively low-cost instruments for research purposes. In the most common TD-NMR implementation [110], static magnetic fields (B_0) are generated by electromagnets or permanent magnets, consisting of a yoke holding two poles at a distance of a few centimeters; they are relatively cheap and can be installed on laboratory benches. B_0 field strengths between 0.12 and 1.4 T, corresponding to proton Larmor frequencies in the range of 5-60 MHz, are obtained. These field strengths gain sufficient sensitivity for observing abundant species in foods such as water and oil. Another approach to non-invasive inspection of intact products is the deployment of single-sided magnets with built-in measurement coils for transmitting and receiving RF signals. A typical TD-NMR instrument with single-sided magnet is shown in Figure 2.7.

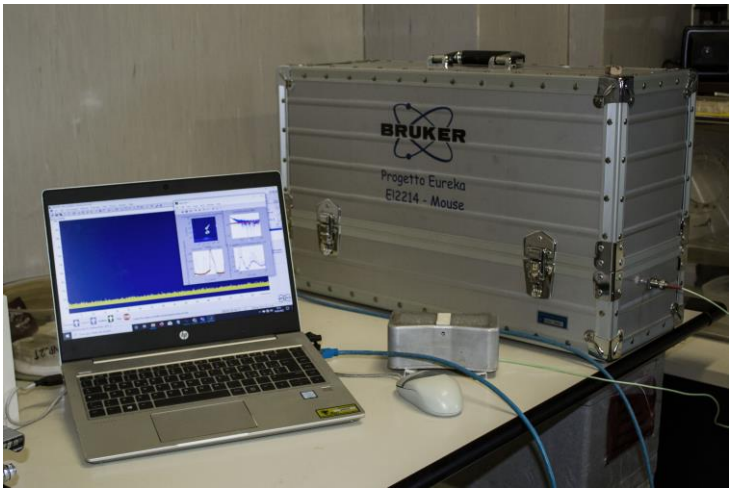


Figure 2.7. TD-NMR single-sided magnet apparatus used in the work described in Chapter 9.

One-dimensional relaxometry involves T_1 and T_2 measurements using the *inversion recovery Fourier transform* (IRFT) method, and the *Carr-Purcell- Meibiim-Gill* (CPMG) pulse sequence, respectively, at a fixed spectrometer frequency and for a single nucleus. The

[110] van Duynhoven, J.; Voda, A.; Witek, M.; Van As, H. In *Annual Reports on NMR Spectroscopy*, Volume 69 (pp. 145-197), Academic Press, **2010**.

advent of a fast algorithm for two-dimensional *Laplace inversion*[111] has led to the development of the so-called multidimensional NMR relaxometry. This transformation can be considered the analogous of the Fourier Transform (time-frequency link for periodic functions) in the establishment of a connection between the time and the time constant for exponential decays. Spin relaxation is indeed manifested as decaying signals; the absence of the periodic character causes a disappearance of some decays during the time of measurement. For this reason, small perturbation of the signal, i.e., noise (caused by machine noise, incorrect or simplified modeling of the system, additional factors or variables that were not included in the model, or varying conditions during different measurements), may lead to a large change in the solution of the fitting.

Since the study of T_1 and T_2 allows characterizing the structure and dynamics of a system, it can be observed that homogeneous systems, such as a solution, show a single couple of (T_1, T_2) relaxation, whereas heterogeneous ones (e.g., food products, biological tissues or inorganic materials[112]) are characterized by the co-existence of different spin-populations which are reflected in a distribution of (T_1, T_2) coupling. The power of the 2D relaxometry is thus best expressed for such kinds of samples, allow investigating the different spin-population even if present in several phases. Discrimination of the various food components is therefore possible investigating differences in relaxation times.

In multidimensional NMR relaxometry, the 2D $f(T_1 - T_2)$ correlation data (Figure 2.8) can be collected using a pulse sequence constituted by two parts (i.e., two different processes governing the spin dynamics) $P(180°) - t - P(90°) - t_e/2 - [P(180°) - t_e]_N$ They are an inversion recovery sequence (first dimension) with a variable recovery time $t_r = Nt_e$ (T_1 measurement) followed by a CPMG sequence (second dimension) acquired in a time t (T_2 measurement). At this point *Laplace inversion* is applied to extract the $f(T_1 - T_2)$ distributions from the measured set of echo amplitudes $A_{T_1,T_2}(t, t_r)$[113]. As stated above, the inversion is ill-conditioned in the sense that small noise in the data can cause large changes in the spectrum and that infinite number of solutions that satisfy the noise model exist. The conventional approaches aim at finding a single solution that is stable against noise selecting a subset of the solutions (e.g., the smooth solutions) and then looking for the best solution within this subset. Commonly used methods include *Tikonov regularization*[114], whose strong implication is introduced in Chapter 9.

---

[111] Song, Y.-Q.; Venkataramanan, L.; Hürlimann, M.D.; Flaum, M.; Frulla, P.; Straley, C. *J. Magn. Reson*. **2002**, *154*, 261-268.

[112] a) Terenzi, C.; Casieri, C.; De Luca, F.; Quaresima, R.; Quarta, G.; Tudisca, V. *Appl. Magn. Reson*. **2015**, *46*, 1159-1178; b) Gabriele, F.; Tortora, M.;Bruno, L.; Casieri, C.; Chiarini, M.; Germani, R.; Spreti, N. J. Cult. Herit. **2021**, *49*, 106-114.

[113] Venkataramanan, L.; Song, Y.-Q.; Hürlimann, M.D. *IEEE Trans. Signal Process*. **2002**, *50*, 1017-1026.

[114] Tikhonov, A.N.; Arsenin, V.Y. New York: Wiley, **1977**.

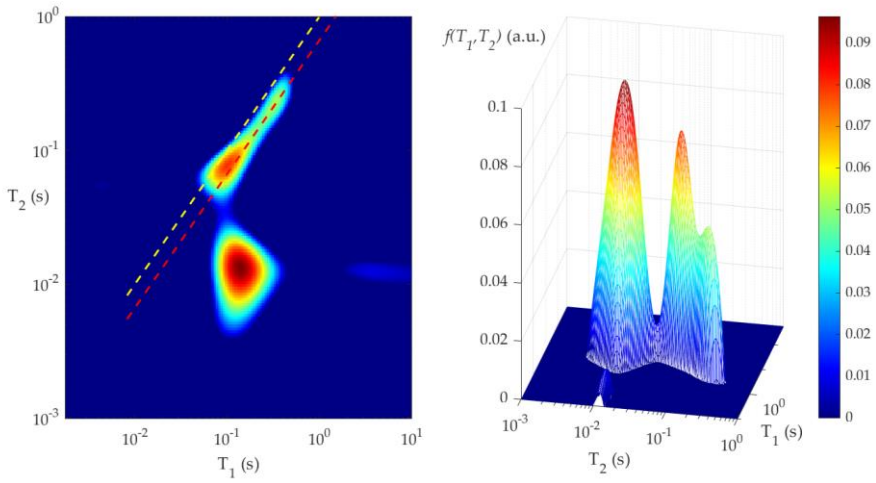


Figure 2.8. 2D ^{1}H NMR T1–T2 distribution function (left) measured on a dairy product (Pecorino cheese). The dashed yellow line corresponds to T_1=T_2, whereas the red one indicates the quasi-diamagnetic reference line T_1=1.5T_2. The 3D-view where $f(T_1, T_2)$ is expressed in arbitrary units (a.u.) is shown on the right. A colorbar is associated depending on the intensity, an internal (to the algorithm) or external normalization of the signals for the maximum intensity can be applied. The region T_1<T_2 of the map is always physically inaccessible for the above-mentioned reasons.

Abbreviations

A

AES, Atomic Emission Spectrometry

AOTF, Acousto-Optic Tunable Filters

ATR, Attenuated Total Reflectance

C

CCD, Charge-Coupled Device

CID, Charge-Injection Device

CPMG, Carr-Purcell-Meibiim-Gill sequence

CTD, Charge-Transfer Detectors

D

DI-SPME, Direct Immersion-Solid Phase MicroExtraction

E

ECD, Electron Capture Detector

F

FID, Flame Ionization Detector

FIR, Far Infrared

FT, Fourier Transform

FTIR, Fourier Transformed Infrared

G

GC, Gas Chromatography

H

HATR, Horizontal ATR

HPLC, High Performance Liquid Chromatography

HSI, Hyperspectral Imaging

HS-SPME, HeadSpace-Solid Phase MicroExtraction

I

ICP, Inductively Coupled Plasma

IR, InfraRed

IRFT, Inversion Recovery Fourier Transform

ISC, International Symposium on Chromatography

IUPAC, International Union of Pure and Applied Chemistry

L

LLE, Liquid–Liquid Extraction

LOD, Limit of Detection

M

MIR, Mid InfraRed spectroscopy

MRI, Magnetic Resonance Imaging

MS, Mass Spectrometry

N

NIR, Near InfraRed spectroscopy

NMR, Nuclear Magnetic Resonance

O

OES, Optical Emission Spectrometry

P

PDA, PhotoDiode Array

R

R, Reflectance

S

SPE, Solid-Phase Extraction

SPME, Solid Phase MicroExtraction

T

T, Transmittance;

TD, Time-Domain;

2D ^{1}H NMR-R, Two-Dimensional Proton Nuclear Magnetic Resonance Relaxometry

U

UHPLC, Ultra-High Performance Liquid Chromatography

UV-Vis, UV-Visible spectroscopy

Chapter 3

Someone could ask: why Chemometrics?

3.1 Introduction

The word *Chemometrics* was introduced in 1972 by Svante Wold and Bruce R. Kowalski, co-founders of the *International Chemometrics Society* in 1974. Thanks to the teachings and proselytism works of the first chemometricians, to the availability of personal computers, and to the utility of this discipline in solving real problems, chemometrics became what it is nowadays, a recognized important science in the Chemistry family.

Since then, a lot of definitions of Chemometrics have seen the light. *Chemometrics and Intelligent Laboratory Systems*, the leading journal in the field, defines it as "*the chemical discipline that uses mathematical, statistical, and other methods employing formal logic to design or select optimal measurement procedures and experiments, and to provide maximum relevant chemical information by analyzing chemical data*".

Figure 3.1 shows in this regard the interaction of Chemometrics with Chemistry, Statistics, Mathematics, Electronics, and Computer Sciences.

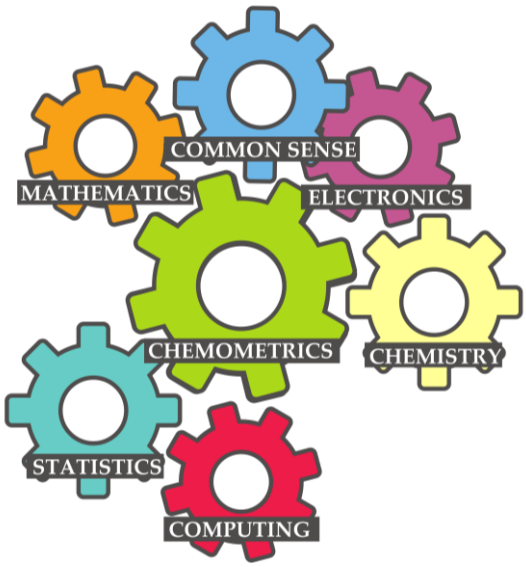


Figure 3.1. Interaction among Chemometrics and the other disciplines. Adapted from Ref. 115.

The advancement of these latter has allowed a constant growth of Chemometrics, expanding the applications of this discipline in all sub-areas of chemistry. In this, it can also be seen *common sense*, reporting the intellectual participation of the operator in the decision-making process during the development of the chemometric studies[115]. This intellectual involvement is indispensable because several actions are established by the operator, including the choice of the variables and their experimental domains, the critical evaluation of the results, and so on.

Chemometricians studied food problems since the origin of Chemometrics. Today food-related issues represent one of the largest fields of application of chemometrics, and food chemical data are used for the development and evaluation of new chemometric techniques and strategies.

The relationship between Chemometrics and Food Chemistry can be obtained by a literature search limited to the past 10 years (2012-2021) through Scopus database, using "Food" AND "Chemometrics" as search terms. About 2900 papers dealing with the application of chemometric tools to food samples have been published and the trend is constantly growing, as it can be seen from Figure 3.2.

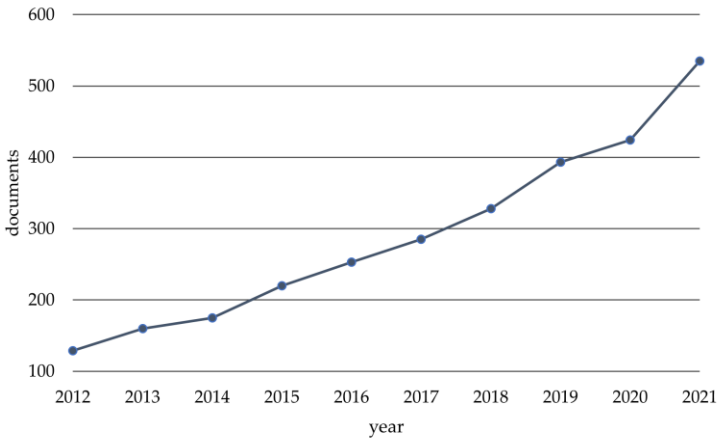


Figure 3.2. Evolution of the number of publications concerning "Food" AND "Chemometrics" retrieved from Scopus database considering the years 2012-2021.

[115] Ferreira, S.L.C. In Worsfold, P.; Townshend, A.; Poole, C. and Mirò, M. (Eds.) *Encyclopedia of Analytical Science* (pp. 420-424), Amsterdam, The Netherlands: Elsevier, **2019**.

Of these, the 82.2% are articles, followed by reviews (10.3%), book chapters (3.3%), and conference papers (2.8%). This is mostly due to the potentiality of the synergistic use of instrumental analytical techniques and chemometrics to obtain trustworthy results in the development of models to solve food-related problems.

Someone could ask: why chemometrics?[116]

In the food area, the system complexity to be faced is increasing, together with the targets of regulatory authorities and society. Food production is connected to environmental, socio-economic challenges, health, safety and nutritional attitudes. Among the emerging research areas there are the study and making of functional food, the monitoring and improvement of the quality of food processing, and the assessment of food authenticity, provenance included. Authenticity and traceability, then, are inherently multivariate, covering many different aspects: the chemical and physical characterization, the adulteration, the discrimination, the mislabeling, and the monitoring of the production process. From the point of view of food and food-processing, this implies that we need to extract information and obtain models capable of inferring the underlying relationships that link the compositional profile and the processing conditions to very general end properties of foodstuff, such as the healthiness, consumer perception, the link to a specific territory and so on. Moreover, the use of different analytical strategies to assess the complex food composition, fundamental to solve problems of identification, process, and quality control, is usually carried out. Prodigious volumes of analytical data are thus produced; unfortunately, not always a lot of data allow obtaining useful information. Chemometrics is necessary in this scenario to obtain relevant information from rough measured data, which, subsequently, improves the interpretation and the presentation of the results.

Generally, Chemometrics includes two macro-families of statistical techniques dealing with issues upstream (*Design of Experiments*, DoE) and downstream (*Multivariate Analysis*) the investigation. This latter embraces D*ata-Mining* (unsupervised and exploratory analysis) and *Machine Learning* (supervised and predictive models, e.g., regression and classification)[117].

General aspects and major implications of the chemometrics approaches used in this thesis will be explored through the following paragraphs.

The data obtained from most of the employed analytical techniques are usually arranged in a matrix $\boldsymbol{X}$ of dimensions NxM, where N is the number of samples (observations) and M that of the experimental variables, generally numerous and highly correlated. Usually,

---

[116] Bertacchini, L.; Cocchi, M.; Li Vigni, M.; Marchetti, A.; Salvatore, E.; Sighinolfi, S.; Silvestri M.; Durante C. In Marini, F. (Ed.) *Chemometrics in Food Chemistry-Data Handling in Science and Technology*, Volume 28 (pp. 371-410), Amsterdam, The Netherlands: Elsevier, **2013**.
[117] Amigo, J.M. *Braz. J. Anal. Chem*. **2021**, *8(32)*, 45-61.

the data are collected together with useful information or *labels* already known. Chemometric elaboration is performed on these datasets.

The following nomenclature and abbreviations are applied in the Chapter.

Nomenclature and abbreviations	
paragraph 3.2. Exploratory Data Analysis	
$\boldsymbol{X}$	generic data matrix (NxM) *i*=1:N for samples (rows), j=1:M for variables (columns)
$\boldsymbol{x}_{ij}$	generic element for matrix $\boldsymbol{X}$
$\boldsymbol{X}_c$	mean-centered matrix, $\boldsymbol{X}_c = \boldsymbol{X} - \boldsymbol{1m}^T$
$\boldsymbol{T}$	scores matrix (NxA) from the decomposition of $\boldsymbol{X}$ in PCA, SCA and PLS decomposition a=1:M for retained components
$\boldsymbol{t}_i$	vector of $\boldsymbol{T}$ collecting the scores for sample *i*
$\boldsymbol{t}_a$	vector of the $\boldsymbol{T}$ describing the projection of the objects onto the ath component
$\boldsymbol{P}$	loadings matrix (MxA) from the decomposition of $\boldsymbol{X}$ in PCA, SCA and PLS decomposition
$\boldsymbol{p}_i$	vector of $\boldsymbol{P}$ collecting the loadings for sample *i*
$\boldsymbol{p}_a$	vector of $\boldsymbol{P}$ collecting the weights of the linear combination $\boldsymbol{t}_a = \boldsymbol{Xp}_a$
$\boldsymbol{E}$	matrix of residuals (NxM)
$\boldsymbol{e}_{ij}$	generic element for matrix $\boldsymbol{E}$, random error
$\boldsymbol{\lambda}$	diagonal matrix of eigenvalues (AxA)
T_i^2	*sample i-PCA model* Mahalanobis distance, i.e., Hotelling statistic
$\boldsymbol{t}_{con,i}$	variables contribution to define T_i^2
Q	*sample i-PCA model* squared orthogonal Euclidean distance, i.e., Q residuals statistic
$\boldsymbol{Q}_{con,i}$	variables contribution to define Q
paragraph 3.3. Coupling of EDA and experimental design: from ANOVA to ASCA	
H_0	null hypothesis
μ	mean of a population, overall offset (estimated through $x_{..}$)
SSQ_b	*sum of squared differences between*-groups
SSQ_w	*sum of squared differences within*-groups
$x_{.g}$	mean for the gth (1:G) group
x_{ig}	ith observation in the gth group (1:N_g)
$x_{..}$	*grand mean*, i.e., total mean value
MSD_b	mean squares deviations *between*-groups
MSD_w	mean squares deviations *within*-groups
x_{fdk}	measured response, replicate k
α_f	effect of the factor α at level f
β_d	effect of the factor βat level d
$\alpha\beta_{fd}$	interaction term between α at level f and β at level d
ε_{fdk}	random error, estimable only if replicates are available
$x_{.f}$	mean for the observations at level f of factor α
$x_{.d}$	mean for the observations at level d of factor β
x_{fd}	average response in the cell fd
$\boldsymbol{X}_\mu, \boldsymbol{b}_\mu$	effect matrix and coefficients accounting for μ
$\boldsymbol{X}_\alpha, \boldsymbol{b}_\alpha$	effect matrix and coefficients accounting for α
$\boldsymbol{X}_\beta, \boldsymbol{b}_\beta$	effect matrix and coefficients accounting for β
$\boldsymbol{X}_{\alpha\beta}, \boldsymbol{b}_{\alpha\beta}$	effect matrix and coefficients accounting for αβ
$\boldsymbol{X}_e$	effect matrix accounting for the residuals
$\boldsymbol{X}^T\boldsymbol{X}$	cross-product, variance-covariance matrix
$\boldsymbol{X}_m$	matrix (NxM) collecting the mean of the M variables (all rows are exactly the same)

$\boldsymbol{m}$	vector of sample means (estimate of $\boldsymbol{\mu}$)
$\boldsymbol{W}$	*within*-group covariance matrix
	paragraph 3.4. Predictive models
$p(\boldsymbol{x}_i \mid g)$	probability that *i-th* observation $\boldsymbol{x}_i$ belongs to a class g in LDA framework
$\boldsymbol{S}$	pooled variance/covariance matrix in LDA framework
p_{0g}	a-priori probability for class g in LDA framework
$\bar{\boldsymbol{x}}_g$	mean (centroid) of class G in LDA framework
$\bar{\bar{\boldsymbol{x}}}$	grand-mean among the classes in LDA framework
$\boldsymbol{S}_b$	the between-group variance/covariance matrix in LDA framework
λ_i	eigenvalue associated to the matrix $\boldsymbol{S}^{-1}\boldsymbol{S}_b$ in LDA framework
Λ_j	univariate Wilk's Lambda $\frac{SSQ_w}{SSQ_{tot}}$ for the variable j in LDA framework
SSQ_{tot}	total sum of squares $SSQ_b + SSQ_w$ in LDA framework
$\boldsymbol{R}$	cross-product matrix in LDA framework
Λ	multivariate or global Wilk's Lambda $\frac{\det(R_w)}{\det(R)}$ in LDA framework
$\boldsymbol{Y}$	experimental response matrix (NxG) g=1:G where G are classes in a classification problem or properties in a regression one g is here used in general to indicate groups (like treatments in ANOVA)
$\boldsymbol{y}_{ig}$	generic element for matrix $\boldsymbol{Y}$
$\hat{\boldsymbol{Y}}$	calculated response matrix (NxG)
$\boldsymbol{B}$	matrix of regression coefficients (MxG)
$\boldsymbol{b}$	vector of matrix $\boldsymbol{B}$
$\boldsymbol{X}_{new}$	novel set of data, not used in calibration
$\hat{\boldsymbol{Y}}_{new}$	predicted response matrix for $\boldsymbol{X}_{new}$
$\boldsymbol{U}$	scores matrix (NxA) from the decomposition of $\boldsymbol{Y}$
$\boldsymbol{Q}$	loadings matrix (GxA) from the decomposition of $\boldsymbol{Y}$
$\boldsymbol{E}_Y$	matrix of residuals (NxG) from the decomposition of $\boldsymbol{Y}$
CE_{CV}	classification error in cross-validation
vip_{m}	variable importance in projection for the *m-th* variable
$\mathrm{w}^2_{\mathrm{ma}}$	squared PLS-weight of the mth variable for the a^{th}component
$\mathrm{SSY}_{\mathrm{a}}$	variance in the response explained by the a^{th}component
$\mathrm{SSY}_{\mathrm{tot}}$	total variance of the response explained by the model
$\boldsymbol{Z}$	general 2nd data-block of dimensions (N×M_2) in MB-framework
$\boldsymbol{C}$	regression coefficients for $\boldsymbol{Z}$ (M_2xG)
$\boldsymbol{Z}_{orth}$	matrix produced by the orthogonalization of $\boldsymbol{Z}$ with respect to $\boldsymbol{T}_X$ score matrix, derived by $\boldsymbol{X}$ (NxM)
$\boldsymbol{Z}_{proj}$	matrix produced by the projection of $\boldsymbol{Z}$ with respect to $\boldsymbol{T}_X$ score matrix

3.2 Exploratory Data Analysis (EDA)

In the scenario described in the Introduction, the *inductive* data-driven research approach with a minimum of *a-priori* knowledge about the objects under exam, rather than the *deductive* one, is an effective way to proceed. This means, figuratively, that it is not the operator asking questions to the data; the data, instead, are allowed talking, giving evidence of their relationships, similarities, differences, trends, clusters and correlations. EDA, through a data-compression, works revealing hidden information from data in such a form that the operator obtains immediate, direct, and *easy-to-understand* representations (graphical outputs).

Principal Component Analysis[118] (PCA) is one of the most well-known exploratory analysis methods, widely applied in different research fields. PCA allows representing multivariate data in a low-dimensionality space of mutually orthogonal, thus uncorrelated, *principal components* (PCs), defined by linear combinations of original variables explaining unrelated portions of information.

Its formulation is based on the bilinear decomposition of the data matrix $\boldsymbol{X}$ (NxM) described in eq. 3.1:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T \qquad \text{(eq. 3.1)}$$

where $\boldsymbol{T}$ (with dimension NxA) is the scores matrix, expressing the coordinates of the samples in the PC space, and $\boldsymbol{P}$ is the loadings matrix (MxA) that defines the new directions, by weighing the relationship between the original variables and the PCs. The dimension A can assume a value from 1 to M. When A=M the PCA model is expressed by eq. 3.1, whereas eq. 3.2 describes the PCA modelling when A<M. More in details, a residual matrix $\boldsymbol{E}$ (NxM) is introduced to take into account the residuals associated with the approximation of the original data with fewer PCs than the original variables[119]. Normally, there is no need of using all the scores and loadings to represent the data, hence, $\boldsymbol{T}$ and $\boldsymbol{P}^T$ are truncated:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \qquad \text{(eq. 3.2)}$$

where $\boldsymbol{E}$ is the residuals matrix.

---

[118] a) Pearson, K. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559-572; b) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37-52.

[119] Oliveri, P.; Forina, M. In Picò, Y. (Ed.) *Chemical Analysis of Food: Techniques and Applications* (pp. 25-57), Amsterdam, The Netherlands: Elsevier, **2012**.

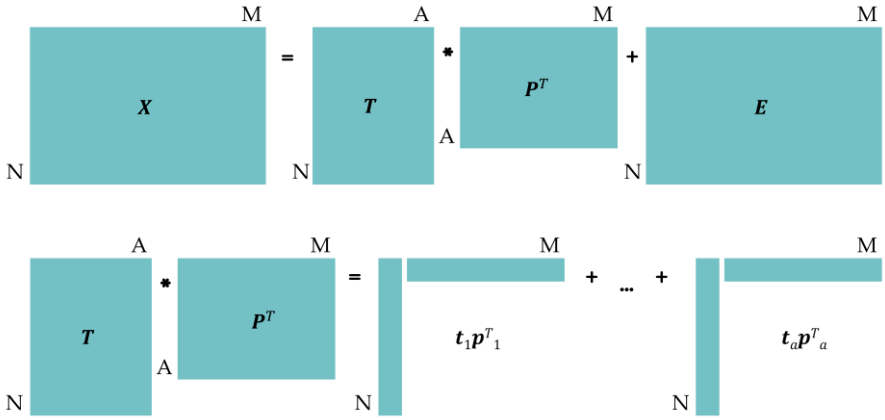


Figure 3.3. Schematic representation of the PCA-decomposition of the matrix $\boldsymbol{X}$.

Figure 3.3 outlines the nature of *PCs as linear combinations of the original variables*. From the geometrical standpoint, PCA can be seen as a projection method, whose aim is to project the observations from the space of the original variables to the PCs-subspace (scores plot). This operation is particularly convenient because, by construction, PCs represent the directions which allow maximizing the variance. Moreover, principal components are iteratively extracted, under the orthogonality constrain (i.e., each PC is orthogonal with respect to all the previous extracted PCs) providing the most informative, but, at the same time, most reduced, set of latent variables. This opens to a number of benefits, e.g., to the possibility of realizing graphical representations which allow detecting outliers or inspecting specific trends/grouping tendencies in data[120].

PCA score plot, indeed, allows providing a graphical and straightforward visualization of the trends within the data samples; moreover, the interpretation of the selected (it is informative enough to visualize the data distribution by considering just the first components) PCs in terms of the original variables is possible through the loading plot.

$\boldsymbol{t}_a$ is the vector of the scores, which describes the projection of the objects onto the ath component, and $\boldsymbol{p}_a$ is the loading vector, which collects the weights of the linear combination in the eq. 3.3:

$$\boldsymbol{t}_a = \boldsymbol{X}\boldsymbol{p}_a \qquad \text{(eq. 3.3)}$$

Since the matrix $\boldsymbol{X}$ brings variation relevant to the problem (i.e., information), it seems reasonable to have as much as possible of that variation in $\boldsymbol{t}_a$ also. The variation, measured by the variance ($var(\boldsymbol{t}_a)$), can be maximized by choosing optimal weights for the linear combination ($\boldsymbol{p}_a$). Therefore, the problem can be formulated as a mathematical

[120] Li Vigni, M.; Durante, C.; Cocchi, M. In Marini, F. (Ed.) *Chemometrics in Food Chemistry-Data Handling in Science and Technology*, Volume 28 (pp. 55-126), Amsterdam, The Netherlands: Elsevier, **2013**.

maximization of $var(\boldsymbol{t}_a)$ with two constraints. Since multiplying an optimal $\boldsymbol{p}_a$ with an arbitrarily large number will make the variance of $\boldsymbol{t}_a$ also arbitrarily large, the weights (the loadings vectors) have to be normalized (first constrain: $\boldsymbol{p}_a^T\boldsymbol{p}_a = 1$). To explain the uncorrelated portion of variance, orthogonalization has also to be performed (second constrain: $\boldsymbol{p}_a^T\boldsymbol{p}_{a+1} = 0$)[121].

Two different algorithms are mainly employed in PC model computation: Nonlinear Iterative Partial Least Squares[122] (NIPALS), proposed by Herman Wold in 1966, and Single Value Decomposition[123] (SVD). NIPALS extracts the principal components sequentially, iteratively optimizing the weights until the convergence criterion is met and can be interrupted at any number of factors. It is, therefore, very efficient when only the first few components are required. SVD is, instead, a non-sequential method in which all factors are obtained at the same time. SVD describes data as orthogonal vectors, therefore the information results linearly independent. Another important algorithm is the eigenvalue decomposition[124] (EVD) It differs for the matrix to work on, that is $\boldsymbol{X}$ for SVD and NIPALS and $\boldsymbol{X}^T\boldsymbol{X}$ for EVD.

When the aim is understanding a phenomenon, variability is regarded as an opportunity to capture the information retained in the samples and PCA is a powerful technique to highlight it. In a quality control approach, PCA can also be used as *diagnostic tool* for purposes such as outlier detection and process control. The PCA model can be applied on new process data/samples to detect abnormal behavior due to changes in the system. The *Q-statistic* and the *Hotelling's T^2 statistic* are the two well-known statistics commonly used to this aim. These methods are based on the assumption (generally stemming from the *central limit theorem*) of a normal distribution of the scores. For applying such a statistic, it is necessary to carefully choose a proper number of PCs for the model, just like when used in predictive modelling[102]. Different heuristic criteria for rank analysis have been proposed:

- *a-priori* choosing a minimum amount of variance to be explained (e.g., for class modelling purposes);
- employing the Kaiser's rule, i.e., only the components associated to a variance larger than 1 (the value can be changed to be less restrictive) are retained;
- scree plot, i.e., the plot of the explained/cumulative variance versus the number of principal components[125]. It is the most used graphical method to decide PC-space optimal dimensionality

---

[121] Bro, R.; Smilde, A.K. *Anal. Methods*. **2014**, *6*, 2812-2831.
[122] Wold, H. In Krishnajah, P.R. (Ed.) *Multivariate analysis* (pp. 391-420), NewYork: Academic Press, **1966**.
[123] Golub, G.H.; Reinsch, C. In Bauer, F.L. (Ed.) *Handbook for Automatic Computation. Volume II: Linear Algebra* (pp. 134-151), Berlin, Heidelberg: Springer, **1971**.
[124] Miyashita, Y.; Itozawa, T.; Katsumi, H.; Sasaki, S.-I. *J. Chemom*. **1990**, *4*, 97-100.
[125] Cattell, R.B. *Multivariate Behav. Res*. **1966**, *1*, 245-276.

- through double *cross-validation* (see Appendix B), based on the calculation of the total sum of squared prediction errors between the reconstructed and the original data; components have to be retained until this index is lower than 1

Figure 3.4 outlines the graphical representation of the geometrical means of the *sum of normalized squared scores* T^2 and the *sum of squared residuals* for each sample Q.

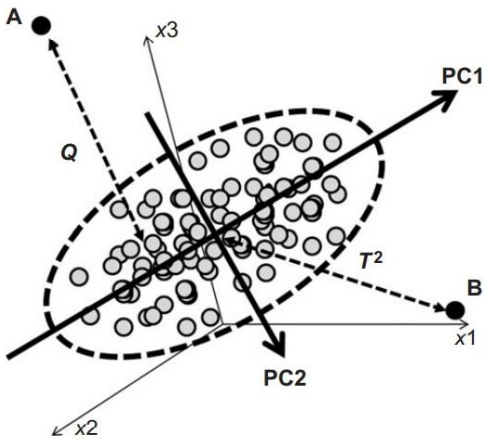


Figure 3.4. Graphical representation of T^2 and Q outliers for a two PCs-model on a three variable data set. Reproduced from Ref 120.

The *Hotelling's* T^2 *statistic*[126] is a measure of the variation of each sample within the model, i.e., the Mahalanobis distance of point B in Figure 3.4. It is defined for the ith observation as:

$$T_i^2 = \boldsymbol{t}_i \lambda^{-1} \boldsymbol{t}_i^T \qquad \text{(eq. 3.4)}$$

where $\boldsymbol{t}_i$ is the ith row of the scores matrix $\boldsymbol{T}$ and λ is a diagonal matrix (AxA) containing the eigenvalues corresponding to the A retained PCs (eigenvectors).

T^2 contribution, i.e., how individual variables contribute to define the Hotelling's T^2 value for the sample i, is calculated as:

$$\boldsymbol{t}_{con,i} = \boldsymbol{t}_i \lambda^{-1} \boldsymbol{P}^T \qquad \text{(eq. 3.5)}$$

[126] Hotelling, H. *Br. J. Math. Stat. Psychol.* **1957**, *10(2)*, 69-79.

The Q statistic is instead a measure of how well each sample fit to the PCA model, i.e., the squared orthogonal Euclidean distance between the sample (point A, Figure 3.4) and its projection into the PCA model. For instance, for $\boldsymbol{x}_i$, i.e. the ith sample in $\boldsymbol{X}$, it is calculated from $\boldsymbol{e}_i$, i.e. the ith row in $\boldsymbol{E}$:

$$Q_i = \boldsymbol{e}_i \cdot \boldsymbol{e}_i^T \qquad \text{(eq. 3.6)}$$

The related Q contribution ($Q_{con,i}$), as well as for T^2, is represented by the squared elements along the ith row of the residual matrix:

$$\boldsymbol{Q}_{con,i} = \boldsymbol{e}_i^2 \qquad \text{(eq.3.7)}$$

From an explorative point of view, the T^2 *versus* Q plot allows the investigation of peculiar samples, and the inspection of the contribution plot can help in relating the detected anomalies to the behavior of specific measured variables.

Often, a PCA performed on the raw data is not very meaningful. In regression analysis, an offset is included and the deviation from such an offset is considered as the interesting variation. To translate this concept to EDA, it is necessary to *mean-center* data. Sometimes it is also necessary to think about the scales of the data, to make the variables more comparable. The importance of data-management before data analysis and the most important pre-processing techniques are discussed in Appendix A.

3.3 Coupling of EDA and experimental design: from ANOVA to ASCA

As stated in the previous paragraph, exploratory analysis is an unsupervised technique thanks to which it is possible to identify groups of samples with high/low similarity, without having *a-priori* information about them. It necessarily implies that it is not considered the possible existence of a design behind the experiment. To circumvent this obstacle, ANOVA-Simultaneous Component Analysis (ASCA) was developed. Before explaining ASCA, a brief digression about ANOVA (ANalysis Of VAriance) is presented below.

3.3.1 ANalysis Of VAriance (ANOVA)

ANalysis Of VAriance[127] (ANOVA) is a univariate statistical test aiming at determining if two or more populations have equal means. To achieve this goal, ANOVA exploits the F-distribution to decide whether the *between-* and *within*-variance (Figure 3.5) are statistically different for the considered groups.

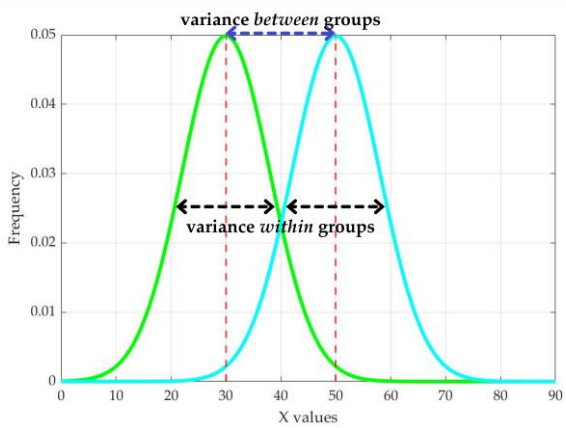


Figure 3.5. Graphical representation of the rationale behind the analysis of variance.

The starting assumption of the test is the null hypothesis (H_0), i.e., the group means are equal except for random errors. The model describing H_0 can be written as in eq. 3.8:

$$x_{ig} = \mu + e_{ig} \qquad \text{(eq. 3.8)}$$

where x_{ig} is the i^{th}observation of the gth group, e_{ig} is the random error, and μ is the mean of the populations, assuming them equal. If the H_0 is not verified, the mean of each gth group has to be assumed: $x_{ig} = \mu_g + e_{ig}$ (eq. 3.9).

Thus, measures of the *between* and *within* group dispersions are needed. Because variances of independent experiments are additive, the squared difference gives a practical measure of dispersion in ANOVA. Thus, the *sum of squared differences between-* (SSQ_b, eq. 3.10) and *within-* (SSQ_w, eq. 3.11) groups are compared[127a]:

$$SSQ_b = \sum_{g=1}^{G} \sum_{i=1}^{N_g} (x_{.g} - x_{..})^2 \qquad \text{(eq. 3.10)}$$

---

[127] a) Ståhle, L; Wold, S. *Chemom. Intell. Lab. Syst*. **1989**, *6*, 259-272; b) Molugaram, K.; Rao, G.S. In Molugaram, K. and Rao, G.S. (Eds.) *Statistical Techniques for Transportation Engineering* (pp. 451-462), Amsterdam, The Netherlands: Elsevier, **2017**.

$$SSQ_w = \sum_{g=1}^{G} \sum_{i=1}^{N_g} (x_{ig} - x_{.g})^2 \qquad \text{(eq. 3.11)}$$

where G and N_g are the number of groups and the number of objects in the gth group, $x_{.g}$ is the mean value of the gth group, $x_{..}$ is the total mean value, and x_{ig} is the ith observation in the gth group.

SSQ_b and SSQ_w are not directly comparable because they should be of the same size; therefore, the number of degrees of freedom *between* and *within* the groups are used as divisors. The variance for the model in the eq. 3.10 is expressed as $\frac{SSQ_b}{(G-1)}$, and the variance for the model in eq. 3.11 is $\frac{SSQ_w}{\sum(N_g-1)}$. These ratios are often referred to as *mean squares deviations* (MSD_b and MSD_w, respectively).

The ANOVA F-test is constructed so that a value of the F-statistic that is sufficiently larger than 1 indicates a difference between the groups. This is accomplished by considering the eq. 3.12 as formula for F[127b].

$$F = \frac{MSD_b}{MSD_w} \text{ if } MSD_b > MSD_w, \text{ or } \textit{vice-versa}: F = \frac{MSD_w}{MSD_b} \text{ if } MSD_w > MSD_b \qquad \text{(eq. 3.12)}$$

Based on the tables of the F-distribution, the F critical value is fixed at a certain p-value (usually 0.05) for the degree of freedom $(G-1)$ and $\sum(N_j - 1)$, hence if $F \geq F_{0.05}$, the differences between the means are significant.

3.3.2 ANOVA applied to designed experiments

When dealing with designed experiments controlling two *factors* (defined as *classifications* which are assumed to be a source of the variability in the data) α and β, considering respectively $f = 1{:}F$ and $d = 1{:}D$ *levels* (defined as *individual classes* within each *classification*), and assuming that at least two measurements are available per *cell* k (i.e., subset of data occurring at the intersection of one *level* of every *factor*), a reasonable ANOVA model would be:

$$x_{fdk} = \mu + \alpha_f + \beta_d + (\alpha\beta_{fd}) + \varepsilon_{fdk} \qquad \text{(eq. 3.13)}$$

where x_{fdk} is the measured response, μ represents an overall *offset* (reference value for the response), α_f and β_d represent the effects of the factors α and β at levels f and d, $\alpha\beta_{fd}$ is the interaction term, and ε_{fdk} is the random error (estimable only if replicates are available).

Eq. 3.13 shows a division of variation on several factors. The basic idea of ANOVA is, indeed, the variation can be separated and assigned to relevant factors. ANOVA is capable of estimating the factor effects by splitting the variations in orthogonal and independent parts, through the usual *sum-to-zero* constraints[128]:

$$\begin{cases} \sum_{f=1}^{F} \alpha_f = 0 \\ \sum_{d=1}^{D} \beta_d = 0 \\ \begin{cases} \sum_{f=1}^{F} (\alpha\beta)_{fd} = 0 \; \forall d \\ \sum_{d=1}^{D} (\alpha\beta)_{fd} = 0 \; \forall f \end{cases} \end{cases} \qquad \text{(eq. 3.14)}$$

These constraints, centering each parameter of the model, allow evaluating them as the incremental contributions induced by each factor at each particular level with respect to μ, estimated as the *grand mean* ($x_{..}$):

$$\begin{cases} \alpha_f = x_{.f} - x_{..} \\ \beta_d = x_{.d} - x_{..} \\ (\alpha\beta)_{fd} = x_{fd} - x_{.f} - x_{.d} + x_{..} \end{cases} \qquad \text{(eq. 3.15)}$$

where $x_{.f}$ is the mean for the observations at level f of factor α, $x_{.d}$ is the mean for the observations at level d of factor β, and x_{fd} represents the average response in the cell fd.

The eq. 3.13 is a linear model (without considering ε_{fdk}); an alternative way to formulate the ANOVA model is therefore through the multiple regression:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} \quad (k = 1) \qquad \text{(eq. 3.16)}$$

where $\boldsymbol{y}$ is a column vector which contains the experimental responses, $\boldsymbol{Xb}$ represents their ANOVA approximation, and $\boldsymbol{e}$ is a column vector containing the residuals. More in details, $\boldsymbol{X}$ will be a dummy-coded matrix constituted by sub-matrices encoding for the effect of the grand mean ($\boldsymbol{X}_\mu$), the factors ($\boldsymbol{X}_\alpha$ and $\boldsymbol{X}_\beta$) and their interaction ($\boldsymbol{X}_{\alpha\beta}$):

$$\boldsymbol{X} = [\boldsymbol{X}_\mu | \boldsymbol{X}_\alpha | \boldsymbol{X}_\beta | \boldsymbol{X}_{\alpha\beta}] \qquad \text{(eq. 3.17)}$$

When entering categorical variables (i.e., factors) into a regression model, a coding system, that determines the interpretation of the model coefficients, addressing the different factor levels has to be defined. Two commonly used coding are *reference* (one level is selected as the reference and set to zero, while the other levels are represented by one of the indicator variables taking on the value 1) and *sum coding*[129]. The last one is the most appropriate in the context of ASCA for reasons of simplicity and *sum-to-zero* constraints. In general, each main effect with f levels is coded with $f - 1$ columns. Under the *sum-to-zero* constraints, the first $f - 1$ levels are coded with zeros and ones,

---

[128] Jansen, J.J.; Hoefsloot, H.C.J.; van Der Greef, J.; Timmerman, M.E.; Westerhuis, J.A.; Smilde, A.K. *J. Chemom.* **2005**, *19*, 469-481.

[129] Madssen, T.S.; Giskeødegård, G.F.; Smilde, A.K.; Westerhuis, J.A. *PLOS Comput. Biol.* **2021**, *17*, e1009585.

and the last one level (f) is coded with -1. The columns coding for the interaction are obtained by the row-wise Kronecker product of the columns of corresponding main-effects[130]. $\boldsymbol{X}_\mu$ contains only ones. At the end, $\boldsymbol{X}$ will specify the Design of Experiments in a dummy notation: each row corresponds to one observation where it is specified at which level of the main and interaction effects this observation was made.

The same partition is applied for the regression coefficients $\boldsymbol{b}$, obtained through the least square estimation:

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{y} = [\boldsymbol{b}_\mu|\boldsymbol{b}_\alpha|\boldsymbol{b}_\beta|\boldsymbol{b}_{\alpha\beta}]^T \qquad \text{(eq. 3.18)}$$

where $(.)^+$ indicates the pseudo-inverse of its matrix argument.

When dealing with multivariate data with underlying experimental design, Multivariate Analysis of Variance[131] (MANOVA) should be employed. The model described in eq. 3.13 is still valid, although the terms are now vector instead of scalar.

$$\boldsymbol{x}_{fdk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_f + \boldsymbol{\beta}_d + (\boldsymbol{\alpha\beta}_{fd}) + \boldsymbol{\varepsilon}_{fdk} \qquad \text{(eq. 3.19)}$$

As for ANOVA when the regression is applied, the first step of MANOVA is to partition the data matrix $\boldsymbol{X}$ according to the principles outlined above:

$$\boldsymbol{X} = \boldsymbol{X}_m + \boldsymbol{X}_\alpha + \boldsymbol{X}_\beta + \boldsymbol{X}_{\alpha\beta} + \boldsymbol{X}_e \qquad \text{(eq. 3.20)}$$

where $\boldsymbol{X}$ is now a matrix NxM of the observations and the rows in $\boldsymbol{X}_m$ contain the sample estimates of the mean of the M variables (all rows are exactly the same), i.e., $\boldsymbol{X}_m = \boldsymbol{1}\boldsymbol{m}^T$ where $\boldsymbol{1}$ is a column vector of ones and $\boldsymbol{m}$ is the vector of samples means (estimate of $\boldsymbol{\mu}$). The mean-centered matrix $\boldsymbol{X}_c = \boldsymbol{X} - \boldsymbol{1}\boldsymbol{m}^T$ is hence partitioned in eq. 3.20. $\boldsymbol{X}_\alpha$ and $\boldsymbol{X}_\beta$ are effect matrices with the sample estimates of the level means for factor A and factor B, respectively, as rows, and $\boldsymbol{X}_{\alpha\beta}$ contains estimates of the interaction effect between the two factors. The application of eq. 3.20 effectively partitions the sum of squares of the elements in $\boldsymbol{X}$ into factor-specific sum of squares:

$$\|\boldsymbol{X}\|^2 - \|\boldsymbol{X}_m\|^2 = \|\boldsymbol{X}_c\|^2 = \|\boldsymbol{X}_\alpha\|^2 + \|\boldsymbol{X}_\beta\|^2 + \|\boldsymbol{X}_{\alpha\beta}\|^2 + \|\boldsymbol{X}_e\|^2 \qquad \text{(eq. 3.21)}$$

At the end, a classical F test is conducted to study if the contribution of each design term to the overall variability observed in the response is statistically significant.

Hypothesis tests in MANOVA are based on the *between*- and *within*-group ($\boldsymbol{W}$) covariance matrix, assuming that the objects are independent and that the covariance matrix $\boldsymbol{W}$ for the residuals is the same for all groups. A mathematical requirement is that $\boldsymbol{W}$ is

---

[130] Bertinetto, C.; Engel, J.; Jansen, J. *Anal. Chim. Acta X*. **2020**, *6*, 100061.
[131] Ståhle, L; Wold, S. *Chemom. Intell. Lab. Syst*. **1990**, *9*, 127-141.

invertible. To overcome this requirement, not always satisfied, many methods have been proposed[132]; one of the most popular is ASCA.

3.3.3 ANOVA-Simultaneous Component Analysis (ASCA)

ASCA[128,133] can be seen as a special case of MANOVA, under the assumption that the within-group correlations between the variables are zero[134]. ASCA operates by partitioning the total variance of the experimental data into the individual contributions induced by the effect of the factors under control and their interactions. The first step is indeed to specify the linear model according to which the data matrix $\boldsymbol{X}$ is decomposed (eq. 3.20). Simultaneous Component Analysis[135] (SCA), a generalization of PCA for the case of several populations sharing a common set of measured variables, is thus applied on the effect sub-matrices derived. Therefore, ASCA essentially carries out two variance decomposition steps. In the first one (ANOVA decomposition), the variance in each variable is decomposed into factor-specific terms; in the second one, the variance in each effect matrix is further decomposed such that the largest amount of variance among the level means for a certain factor is explained by the first PC, and in decreasing order by the following ones[136]. Typically, a permutation testing or bootstrap procedure is used to determine the significance of the observed effects[137].

To guarantee an appropriate deeper description of the method, the notation proposed in the original paper[133a] will be used hereinafter.

If data matrix X_i (K_ixM) collecting the same type of M measurements performed on I (i=1:I) objects on K_i occasions is available, a reasonable simultaneous component (SC) model would be:

$$\boldsymbol{X}_i = \boldsymbol{T}_i \boldsymbol{P}^T + \boldsymbol{E}_i \qquad \text{(eq. 3.22)}$$

where P (JxA) represents the common loadings matrix (A: number of SCs) and T_i (K_ixA) contains the scores of the subjects on the ith measurement occasion.

---

[132] Jansen, J.; Engel, J. In Jaumot, J., Bedia, C. and Tauler, R. *Data Analysis for Omic Sciences: Methods and Applications*, Volume 82 (pp. 301-335), *Comprehensive Analytical Chemistry* Handbook, Amsterdam, The Netherlands: Elsevier, **2018**.

[133] Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.J.A.N.; van der Greef, J.; Timmerman, M.E. *Bioinformatics*. **2005**, *21*, 3043-3048.

[134] Engel, J.; Blanchet, L.; Bloemen, B.; van den Heuvel, L.P.; Engelke, U.H.F.; Wevers, R.A.; Buydens, L.M.C. *Anal. Chim. Acta*. **2015**, *899*, 1-12.

[135] Timmerman, M.E.; Kiers, H.A.L. *Psychometrika*. **2003**, *68*, 105-121.

[136] Bertinetto, C.; Engel, J.; Jansen, J. *Anal. Chim. Acta*. **2020**, *6*,100061.

[137] a) Vis, D.J.; Westerhuis, J.A.; Smilde, A.K.; van der Greef, J. *BMC Bioinf*. **2007**, *8(1)*, 1-8; b) Anderson, M.; Braak, C.T. *J. Stat. Comput. Simul*. **2003**, *73(2)*, 85-113; c) Timmerman, M.E.; Kiers, H.A.L.; Smilde, A.K.; Ceulemans, E.; Stouten, J. *Br. J. Math. Stat. Psychol*. **2009**, *62(2)*, 299-318.

The model parameters in eq. 3.22 can be found by solving the standard least squares problem:

$$\min_{\boldsymbol{T}_i,\boldsymbol{P}} \sum_{i=1}^{I} \|\boldsymbol{X}_i - \boldsymbol{T}_i \boldsymbol{P}^T\|^2 \qquad \text{(eq. 3.23)}$$

The ASCA analogon for model 3.19 is:

$$\boldsymbol{X}_{hi_h} = \boldsymbol{1}\boldsymbol{m}^T + \boldsymbol{T}_K \boldsymbol{P}_{A=1}^T + \boldsymbol{T}_{Kh} \boldsymbol{P}_{A=2}^T + \boldsymbol{T}_{Khi_h} \boldsymbol{P}_{A=3}^T + \boldsymbol{E}_{hi_h} \qquad \text{(eq. 3.24)}$$

where $\boldsymbol{T}_K$ (KxA1) is the scores matrix containing the contributions of the factor α expressed on the basis $\boldsymbol{P}_{A=1}^T$ (JxA1); the same legend is applied on the following terms.

By imposing the proper constraints, the different parts of the model 3.24 are orthogonal to each other. These constraints are:

$$\begin{cases} \boldsymbol{1}^T \boldsymbol{T}_K = \boldsymbol{0}^T \\ \sum_{h=1}^{H} \boldsymbol{T}_{kh} = \boldsymbol{0} \\ \sum_{i_h=1}^{I_h} \boldsymbol{T}_{Khi_h} = \boldsymbol{0} \;\; \forall\, h = 1{:}H \end{cases} \qquad \text{(eq. 3.25)}$$

In practice, generalizing, the centered data matrix ($\boldsymbol{X}_C$) is decomposed in the sum of as many contributions as there are factors and interactions (plus the residuals) according to the ANOVA scheme; if the effect of a factor (or an interaction) turns out to be significant, a SCA is performed on the corresponding sub-matrix. To assess the extent of the contribution of each factor, i.e., its significance, the *Frobenius norm* of the corresponding effect matrix is calculated and its comparison with the distribution of the values of the sum of squares under the null hypothesis is performed. The distribution corresponding to the null hypothesis (no significance of the controlled effects) is estimated non parametrically with a permutation test[133a], working by means of a randomization of the group labels in the data. The variable profile for a given sample remains unchanged, but the sample is randomly reassigned to one of the other treatment groups. This procedure is repeated for a proper number of times providing an empirical distribution of the SSQ_i values for the null hypothesis and the validation of the significance of the effect time is carried out comparing it with the real one.

The bootstrap procedure [138] is instead a resampling-based method for obtaining inferential information, used in this contest to estimate the confidence intervals of the loadings. The procedure is applied on the single data matrix $\boldsymbol{X}$ to create many simulated datasets ($\boldsymbol{X}_{bootstrap,i}$) and a SCA model, in particular the loadings matrix ($\boldsymbol{P}_{bootstrap,i}{}^T$), is computed on each of them. From the bootstrap distribution of the statistic(s) of interest, i.e., $\boldsymbol{P}^T$, inferential information, like confidence intervals, is derived.

---

[138] a) Efron, E.; Tibshirani, R.J. *An Intoduction into the bootstrap*. New York: Chapman & Hall. **1993**; b) Davison, A.C.; Hinkley, D.V. *Bootstrap methods and their application*. Cambridge: Cambridge University Press. **1997**.

3.4 Predictive models

Experimental data are often collected with the aim of predicting the value of one or more properties of the system (*responses*, generally collected in the *response matrix* $\boldsymbol{Y}$ or *response vector* $\boldsymbol{y}$), both of quantitative or qualitative nature.

Multivariate Regression (MR) models aim at predicting quantitative properties of the objects under study by approximating them through a function applied to the predictor matrix $\boldsymbol{X}$ by means of proportionality constants collected in a regression coefficients matrix (usually indicated as $\boldsymbol{B}$), starting from the assumption that any variation in $\boldsymbol{X}$ is correlated to the variations in $\boldsymbol{Y}$. MR has many advantages over the *classical univariate regression*, e.g., it is less likely affected by overfitting and reduces data dimensionality by searching for a sub-space in the multivariate space where performing the regression. There are some cases in which not only a single matrix of predictor $\boldsymbol{X}$ is measured and used: that is the case of Multi-Block Regression, and Multi-Block techniques in general, that will be fully described in Section 3.4.3. MR can be both linear and non-linear[139]; among the linear ones, Partial Least Squares (PLS) is the best known. In the present thesis it was mainly used as part of the Partial Least Squares-Discriminant Analysis algorithm in classification problems, only in Chapter 10 an application in regression is shown. For this reason, it will be discussed in detail in Section 3.4.2.

When the response to be predicted is qualitative, it may assume only a discrete set of values (without an ordered relationship among them). A qualitative variable induces a categorization, according to which each of the values it can take is said to be a *class* (or *category*). *Classification* methods are chemometric tools building models aiming at predicting which *class* more accurately describes the individuals under investigation based on the experimental data collected[140]. Independently of the kind of problem under investigation, data splitting is a preliminary phase that has to proceed the actual model building because predictive techniques fall into the category of the so-called *supervised methods*. This concept is fundamental in data analysis, and it is discussed more in details in Appendix B.

---

[139] Westad, F.; Bevilacqua, M.; Marini, F. In Marini, F. (Ed.) *Chemometrics in Food Chemistry-Data Handling in Science and Technology*, Volume 28 (pp. 127-170), Amsterdam, The Netherlands: Elsevier, **2013**.

[140] Cocchi, M.; Biancolillo, A.; Marini, F. In Jaumont, J., Bedia, C. and Tauler R. (Eds.) *Data Analysis for Omic Sciences: Methods and Applications* (pp. 265-299), *Comprehensive Analytical Chemistry*, Amsterdam, The Netherlands: Elsevier, **2018**.

3.4.1 Classification methods

Classification aims at characterizing an object (sample) by assigning it to a category on the basis of the measurements made on it. From a geometric point of view, since each sample can be represented as a point in the multivariate space of variables, the classification carries out the identification of surfaces in the multidimensional space, which define the regions where the samples of a particular category are more likely to be found. Two approaches, schematized in Figure 3.6, can be pursued, i.e., the discriminant and the modelling one. In the first case, the geometric space is simultaneously divided in as many regions as the number of classes in the training set, and an object is assigned to only one class. In the latter case, every class is defined independently from the other ones and a sample could be assigned to one class, more than one class or even not be assigned[102a,141].

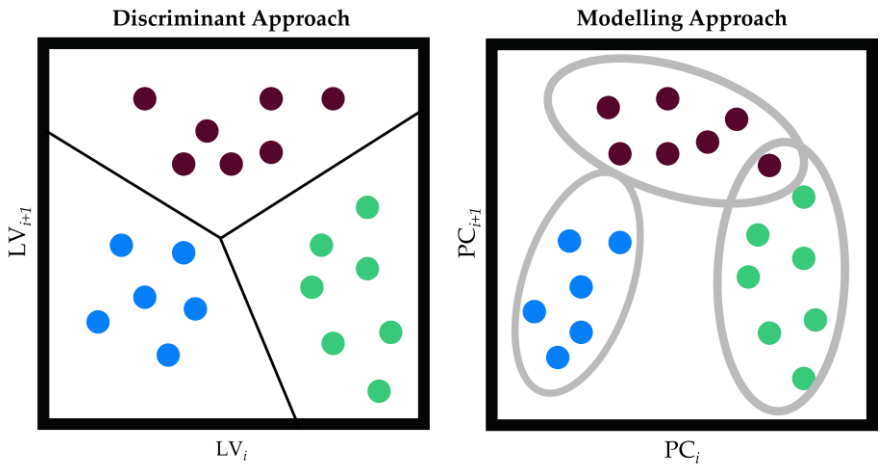


Figure 3.6. Comparison between discriminant and class-modelling approaches.

In the following paragraph, an overview of the discriminant approach, the one employed in this thesis, is provided. Multi-block classification is addressed in Section 3.4.3.

[141] Biancolillo, A.; Marini, F.; Komsta, Ł.; Vander Heyden, Y.; Sherma, J. In Komsta, Ł.; Vander Heyden, Y. and Sherma, J. (Eds.) *Chemometrics in chromatography* (267-284), Boca Raton: CRC Press, **2018**.

3.4.2 Discriminant classification: Linear Discriminant Analysis (LDA) and Partial Least Squares-Discriminant Analysis (PLS-DA)

As stated above, the discriminant approach leads to a univocal classification. The boundaries, defining the class region in the multidimensional space, can be linear or non-linear depending on the selected method. When applicable, linear methods are usually preferred to other approaches (such as Quadratic Discriminant Analysis) because they require the estimation of fewer parameters and have proved to be, in general, more robust against overfitting[142]. *Bayes' theorem* is the standing point for sample assignation in discriminant approaches: the probabilities that a sample belongs to each of the categories represented in the training set (see Appendix B) is computed; successively, each sample is assigned to the class with the highest value of the computed probability to belong to.

Linear Discriminant Analysis (LDA) is one of the first classification methods proposed in the literature. It was conceived by Fisher in 1936[143] and it has paved the way for the entire family of discriminant classifiers. The starting point of LDA is the hypothesis the investigated populations follow a normal distribution; consequently, the likelihood the *i-th* observation $\boldsymbol{x}_i$ belongs to a class g, is defined as:

$$p(\boldsymbol{x}_i|g) = \frac{\Gamma_g}{(2\pi)^{\nu/2}|\boldsymbol{S}|} e^{-0.5(x_i-\bar{x}_g)\boldsymbol{S}^{-1}(x_i-\bar{x}_g)^T} \qquad \text{(eq. 3.26)}$$

where $\boldsymbol{S}$ is the *pooled variance/covariance* matrix, ν the number of the collected variables, and Γ_g ensures each class-probability is 1. In a multi-category case (G>2), the G class-regions are delineated in agreement with eq. 3.27:

$$p(A|\boldsymbol{x}) > p(B|\boldsymbol{x}) \qquad \forall B = 1 \dots G, B \neq A \qquad \text{(eq. 3. 27)}$$

indicating a sample will appertain to Class A when the posterior probability it belongs to this class is higher than the posterior probability it belongs to any other category.

In LDA, the boundaries defined to separate the regions of the multivariate space occupied by different groups, following the Bayes' classification rule, are the linear surfaces where the probability for an object to belong to each of the considered classes is equal[144].

---

[142] Biancolillo, A.; Marini, F. *Front. Chem.* **2018**, *6*, 1-14.
[143] Fisher, R.A. *Ann. Eugen*. **1936**, *7*, 179-188.
[144] Tharwat, A.; Gaber, T.; Ibrahim, A.; Hassanien, A.E. *Al Commun.* **2017**, *30(2)*, 169-190.

The equation for the calculation of these hyperplanes will be calculated as described in eq. 3.28:

$$ln\frac{\Gamma_A p_{0A}}{\Gamma_B p_{0B}} - 0.5(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B) + (\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)^T \boldsymbol{S}^{-1}\boldsymbol{x} = 0 \qquad \text{(eq. 3.28)}$$

$$\forall B = 1 \dots G, b \neq a$$

where p_{0A}, $\bar{\boldsymbol{x}}_A$, p_{0B} and $\bar{\boldsymbol{x}}_B$ are the a-priori probability and the mean of class A and B, respectively.

Data pattern can be visualized by scores and loadings on the *canonical variables* (or *discriminant functions*), which are the directions that maximize the between-class to within-class variance ratio. Analogously to the principal components, discriminant functions are a linear combination of the original variables and, just like in the case of PCA, the computation of the canonical variates is reduced to an eigenvalues/eigenvectors problem[145]. However, while PCA selects directions of the maximum portion of the original variance, LDA identifies directions that maximize the separation between the given classes.

It is possible to show that the canonical variates are the directions identified by $\boldsymbol{w}_i$ (i.e., the weights in the linear combination of original variables) and the largest eigenvalue λ_i of $\boldsymbol{S}^{-1}\boldsymbol{S}_b\boldsymbol{w}_i$. Therefore they are derived solving the eq. 3.29:

$$\boldsymbol{S}^{-1}\boldsymbol{S}_b\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i \qquad \text{(eq. 3.29)}$$

where $\boldsymbol{S}_b$ is the between-group variance/covariance matrix defined as $\boldsymbol{S}_b = \frac{1}{G}\sum_{g=1}^{G}(\bar{\boldsymbol{x}}_g - \bar{\bar{\boldsymbol{x}}})(\bar{\boldsymbol{x}}_g - \bar{\boldsymbol{x}})^T$, $\bar{\boldsymbol{x}}_g$ is the mean of class g and $\bar{\bar{\boldsymbol{x}}}$ is the grand-mean.

Eq. 3.29 reveals the main issue of LDA: the invertibility of $\boldsymbol{S}$. This is a condition rarely met, because it implies the number of observations exceeds the number of collected variables; nevertheless, when this condition is satisfied, the problem can be easily solved by identifying the eigenvalues (λ_i) associated to the matrix $\boldsymbol{S}^{-1}\boldsymbol{S}_b$. In a number of cases, such as in the work described in Chapter 4, the increase of samples to variables ratio is operated by means of a stepwise selection of predictors. In particular, the statistical approach of *Wilk's Lambda* has been adopted. Wilk's Lambda is a distance (*similarity*) approach searching for the best subset of variables satisfying a defined criterion[146] (all the most informative variables are introduced; the best model complexity is achieved).

The first step of the iterative variable selection process is the introduction of the most discriminant factor, which is identified by the *univariate Wilk's Lambda*. Thus, for each variable j, ANOVA is applied ($SSQ_{tot} = SSQ_b + SSQ_w$) and the Wilk's Lamba is

---

[145] Goodall, C. *Technometrics*. **1988**, *30(3)*, 351-352.

[146] el Ouardighi, A.; el Akadi, A.; Aboutajdine, D. In *2007 International Symposium on Computational Intelligence and Intelligent Informatics* (pp. 51-55), **2007**.

computed by the ratio of the within-class sum of squares and the total sum of squares ($\Lambda_j = \frac{SSQ_w}{SSQ_{tot}}$). The variables are more discriminative if their corresponding Wilk's Lambda are small. The following step is based on the sequential introduction of the variables that lead to a decrease of the *multivariate Wilk's lambda*, i.e., more than one variable is involved. Now, the total variability is expressed by the total sum of cross-product matrix $\boldsymbol{R}$ whose objects are calculated as follow[146]:

$$r_{ii'} = \sum_{g=1}^{G} \sum_{n=1}^{n_g} x_{ign} x_{i'gn} - \bar{x}_i \bar{x}_{i'} \qquad \text{(eq. 3.30)}$$

where $r_{ii'}$ is the *i-th* element of the matrix $\boldsymbol{R}$, and G and n_g are the number of classes and the number of objects belonging to the *g-th* class, as usual. Like the univariate analysis of variance, the sum of cross-product matrix $\boldsymbol{R}$ is decomposed into two matrices: the within-class sum of the cross-product matrix ($\boldsymbol{R}_w$), and the between-class sum of the cross-product matrix ($\boldsymbol{R}_b$) For a given set of variables, the *multivariate or global Wilk's Lambda* is computed by the ratio:

$$\Lambda = \frac{\det(R_w)}{\det(R)} \qquad \text{(eq. 3.31)}$$

The more the variables are globally discriminative, the more the Wilk's Lambda is small.

Even if the above approach could be very useful for the variable selection, which is mandatory to apply LDA, the performance of the stepwise LDA could be anyway affected by multicollinear predictors, especially when a multi-class problem with different class dispersions is handled. An alternative to the variable selection can be a feature-extraction method, such as principal components or latent variables (PLS), which take into account the correlation with the response.

Partial Least Squares [147] (PLS) is the most widely used regression method. It allows fitting a response matrix $\boldsymbol{Y}$ to a predictor matrix $\boldsymbol{X}$, even when handling ill-conditioned matrices. The PLS-regression model is built by estimating the regression coefficients $\boldsymbol{B}$:

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{E} = \boldsymbol{XB} + \boldsymbol{E} \qquad \text{(eq. 3.32)}$$

being $\hat{\boldsymbol{Y}}$ the predicted responses and $\boldsymbol{E}$ the residuals.

PLS regression lays on the assumption that both the independent matrix $\boldsymbol{X}$ and the dependent matrix $\boldsymbol{Y}$ can be projected onto a low-dimensional space and that a linear relationship exists between the scores of the two blocks.

---

[147] Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst*. **2001**, *58(2)*, 109-130.

Thus, the two matrices are decomposed into scores ($\boldsymbol{T}$ and $\boldsymbol{U}$) and loadings ($\boldsymbol{P}$ and $\boldsymbol{Q}$) according to the following relations:

$$\begin{cases} \boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}_X \\ \boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{E}_Y \end{cases} \qquad \text{(eq. 3.33)}$$

Moreover, a linear dependence, between the $\boldsymbol{X}$- and $\boldsymbol{Y}$- scores, is assumed (*inner relation*):

$$\begin{cases} \boldsymbol{U} = \boldsymbol{C}\boldsymbol{T} \\ \boldsymbol{Y} = \boldsymbol{C}\boldsymbol{T}\boldsymbol{Q}^T + \boldsymbol{E}_Y \end{cases} \qquad \text{(eq. 3.34)}$$

where $\boldsymbol{C}$ is a diagonal matrix of the coefficients of the inner relation. PLS-based methods allow dealing with highly-correlated variables by searching for the orthogonal directions (i.e., the *latent variables*, LVs) of maximum covariance with the $\boldsymbol{Y}$ while ensuring an explanation of the relevant sources of data variability[148]. A crucial step in all latent variables-based methods) is the selection of the optimal number of LVs to be retained. Generally, this is done by calculating and plotting the Error in Cross-Validation *vs* the increasing number of LVs. In the context of classification, in particular, the Classification Error in Cross-Validation (CE_{CV}) is evaluated:

$$CE_{CV} = \frac{\sum_{i=1}^{N} e_{-i}}{N} \qquad \text{(eq. 3.35)}$$

where the error term e_{-i} indicates the cross-validated error in prediction for the ith sample. More details about cross-validation and prediction procedures will be given in Appendix B, together with a description of the main indicators of models' performances.

The solution of a classification problem can be achieved by PLS when the so-called *dummy matrix* $\boldsymbol{Y}$ is used as a response matrix binary encoding the class-belonging.

Partial Least Squares–Discriminant Analysis[149] (PLS-DA) is the so-derived discriminant classifier. PLS-DA is a flexible and effecting well-known classification method, particularly suitable to handle correlated variables. The method is based on the exploitation of a dummy matrix which cyphers samples' class-membership. This response matrix has as many rows as the number of training samples and as many columns as the number of the investigated categories. Each row-vector will present a 1 in correspondence of the column associated to the actual class of the corresponding sample, and 0*s* in all the other entries. For instance, for a three classes problem (as the

---

[148] Nocairi, H.; Qannari, E.M.; Vigneau, E.; Bertrand, D. In Colubi, A.M.; Kontoghiorghes, E. and Park, B.(Eds.) *Computational Statistics & Data Analysis*, Volume 48(1) (pp. 139-147), Amsterdam, The Netherlands: Elsevier, **2005**.

[149] a) Sjöström, M.; Wold, S.; Söderström, B. In Gelsema, E.S and Kanal L.N. (Eds.) *Pattern Recognition in Practice II* (pp. 461-470), Amsterdam, The Netherlands: Elsevier, **1986**; b) Ståhle, L.; Wold, S. *J. Chemom*. **1987**, *1(3)*, 185-196; c) Barker, M.; Rayens, W. *J. Chemom*. **2003**, *17(3)*, 166-173.

one inspected in Section 1), samples belonging to Class 1, Class 2 and Class 3, respectively, will be represented by the row vectors:

$$\begin{cases} y_1 = [1\ 0\ 0] \\ y_2 = [0\ 1\ 0] \\ y_3 = [0\ 0\ 1] \end{cases} \qquad \text{(eq. 3.36)}$$

Once the response is generated, the general regression equation $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$, already described, can be solved by PLS. In the definition of the model the evolution of CE_{CV} in eq. 3.35 is considered.

Once $\boldsymbol{B}$ has been estimated on the training samples, if a novel set of data ($\boldsymbol{X}_{new}$) is available, it is possible to use the calibration model to predict new samples:

$$\widehat{\boldsymbol{Y}}_{new} = \boldsymbol{X}_{new}\boldsymbol{B} \qquad \text{(eq. 3.37)}$$

Conversely to $\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}}_{new}$, is constituted of continuous values and a classification rule is necessary to assign new samples to the different classes. The simplest approach, and almost never used, is to assign the sample to the category for which the $\boldsymbol{Y}$ predicted value is highest, but this procedure is particularly tricky when the classification problem involves more than two categories. Consequently, different and more elaborated criteria have been proposed. In the present thesis, PLS was used as a feature extraction step for the final LDA model (PLS-LDA), which could be computed both on the LVs scores or on the calculated $\widehat{\boldsymbol{Y}}$[150]. Further, different probabilistic approaches could be applied to the $\widehat{\boldsymbol{Y}}$. Among them, a relevant proposal is the one by Perez and collaborators[151], suitable for the two-classes problems. Two Gaussian distributions (one for the in-class samples and the other for all the not-in-class ones) may be built from the continuous $\boldsymbol{Y}$-values and for each column of the predicted $\boldsymbol{Y}$-block. This allows calculating the probability to observe a $\boldsymbol{Y}$-value for a given sample that belongs to the *g-th* class or that does not belong to it. A probabilistic threshold for each class can be pinpointed at the intersection of the curves that, due to the normalization step, corresponds to an a posteriori probability of 0.5. PLS-DA results are usually expressed in terms of percentage of correct classification for every class, i.e., the percentage of correctly classified samples.

Once the model is computed, variable interpretation can be performed. Variable importance in projection[152] (VIP) is a feature ranking approach suitable for identifying which variables contribute the most to the calculation of a projection-based model.

---

[150] Indahl, U.G.; Martens, H.; Næs, T. *J. Chemom.* **2007**, *21(12)*, 529-536.
[151] Pérez, N.F.; Ferré, J.; Boqué, R. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 122-128.
[152] Wold, S.; Johansson, E.; Cocchi, M. In Kubinyi, H. (Ed.) *3D QSAR in Drug Design, Theory, Methods, and Applications*. Leiden: ESCOM Science Publishers, **1993**.

For instance, given a PLS-DA model, calculated extracting A latent variables, the *m-th* VIP index (related to the *m-th* variable) vip_{m} is calculated as:

$$vip_{\mathrm{m}} = \sqrt{\sum_{\mathrm{a}=1}^{\mathrm{A}} \frac{\mathrm{w}_{\mathrm{ma}}^{2}\mathrm{SSY_aM}}{\mathrm{SSY_{total}A}}} \qquad \text{(eq. 3.38)}$$

where $\mathrm{w}_{\mathrm{ma}}^{2}$ is the squared PLS-weight of the mth variable for the a^{th}component, $\mathrm{SSY_a}$ is the variance in the response explained by the ath component, M the total number of original variables and $\mathrm{SSY_{tot}}$ is the total variance of the response explained by the model. Customarily, the value 1 is considered a convenient threshold to identify the chemical elements contributing the most to model-building; consequently, in the present work, all the variables presenting VIP>1 are identified as highly relevant.

3.4.3 Multi-block (MB) data analysis: an overview

In recent years, *multi-block* data has become widely popular. It can either come from a multi-platform analysis of the same samples, or by the combination of chemical measurements with non-analytical data (e.g., from sensory and consumer sciences), or can be obtained when working under different conditions, i.e., when multiple batches are available from an industrial process performed under diverse circumstances. In all these cases, the data is not simply multivariate but is *multi-modal*, i.e., multi-source. The main characteristics of multi-block data is that, if a common mode is shared (i.e., the samples mode)[153], it can consist either of multiple matrices collecting properties acquired from different instruments on the same samples (Figure 3.7, A), or of a combination of matrices and higher order arrays (Figure 3.7, B), or of different and *independent* (in this case this multi-set analysis falls under the multi-block framework) batch processes (Figure 3.7, C)[154].

---

[153] Lahat, D.; Adali, T.; Jutten, C. *Proc. IEEE Inst. Electr. Electron. Eng.* **2015**, *103(9)*, 1449-1477.
[154] Mishra, P.; Roger, J.-M.; Jouan-Rimbaud-Bouveresse, D.; Biancolillo, A.; Marini, F.; Nordon, F.; Rutledge, D.N. *Trends Analyt. Chem.* **2021**, *137*, 116206.

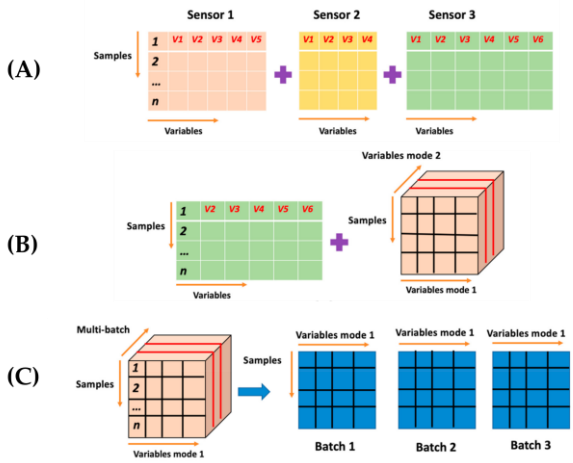


Figure 3.7. Typical multi-block data configurations in analytical chemistry. Data generated by multiple analytical platforms in the form of 2D matrices (A), or in the form of 2D matrices and higher order tensors (B), and multi-set data from batch processes (C). Adapted from Ref. 154.

Traditional *single-block* chemometrics techniques work properly when the data is generated by only one source of variability, whereas they extract only a limited part of the information present in the data when dealing with cases such as those in Figure 3.7. Complementary information from data generated in multiple modes can be extracted exclusively when *multi-block* (or *data fusion*, DF) *data analysis* is performed. MB techniques accomplish the enhancement of data visualization, the improvement of predictive performances, and the identification of key variables influencing the models, when achieving an improvement understanding of the common and distinct information, as well as data-block interactions[155].

Many classifications of DF methods have been proposed in literature. Depending on the level at which data can be combined, it is possible to categorize the strategy as *low-level*, *mid-level* or *high-level*[156] (not performed in this thesis) data fusion. These are represented in Figure 3.8.

---

[155] Borràs, E.; Ferré, J.; Boqué, R.; Mestres, M.; Aceña, L.; Busto, O. *Anal. Chim. Acta*. **2015**, *891*, 1-14.
[156] Ballabio, D.; Todeschini, R.; Consonni, V. In Cocchi, M. (Ed.) *Data Fusion Methodology and Applications* (pp. 129-155), Part of *Data Handling in Science and Technology* Book Series, Volume 31. Elsevier, **2019**.

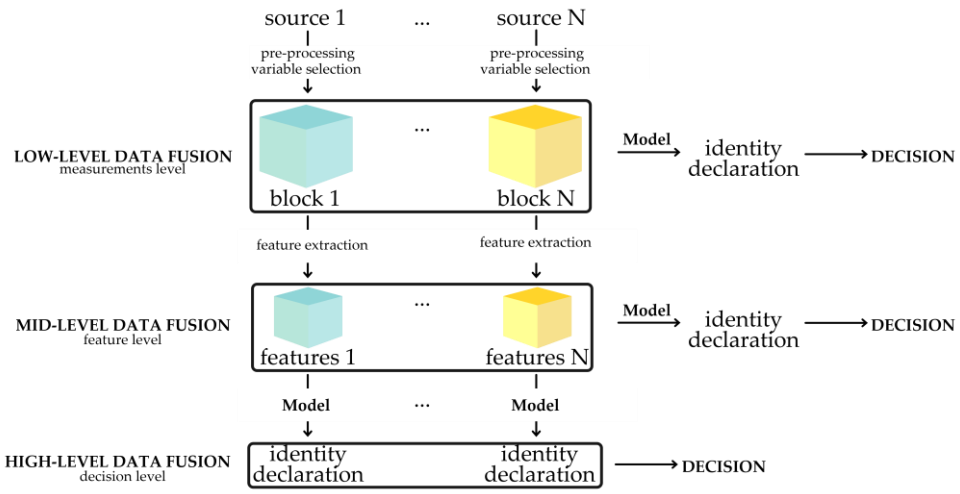


Figure 3.8. Data fusion scheme at low-, mid- and high-levels.

In the low-level DF, the original datasets are concatenated according to the shared mode before any other data elaboration[157]. It is a direct and easy approach that, on the other hand, tends to generate merged matrices potentially redundant at an informative level or unbalanced due to a greater weight of certain techniques with respect to others. The mid-level DF consists of extracting the relevant features from each data-block and of concatenating them into a single matrix, which is further elaborated. Since this strategy works at the features level, information is concentrated in a few variables and most of the noise and irrelevant sources of variability are eliminated. For these reasons, this approach generally provides the most satisfying results. In this case, the feature extraction may be performed independently for each block (mid-level DF), thus the blocks are exchangeable. Usually scores PCs from principal component analysis or scores on LVs from partial least squares, both in regression (PLS) and classification (PLS-DA) are used[158]. The challenge is to find the optimal combination of extracted features that describe the significant variation of the instrumental responses and provides the best final model. The high-level DF, not explored in this thesis, takes place at the so-called *decision level*: for each data-block a model is built and, eventually, the various models are merged to produce the responses.

---

[157] Biancolillo, A.; Boqué, R.; Cocchi, M.; Marini, F. In Cocchi, M (Ed.) *Data Fusion Methodology and Applications* (pp. 271-310), Part of *Data Handling in Science and Technology* Book Series, Volume 31. Elsevier, **2019**.
[158] Biancolillo, A.; Bucci, R.; Magrì, A.L.; Magrì, A.D.; Marini, F. *Anal. Chim. Acta*, **2014**, *820*, 23-31.

Multi-block analysis is influenced, just like the standard one block methods, by the pre-processing operations; therefore, they should be carefully planned in accordance with the multi-block analysis to be performed.

In MB scenario, data pre-treatment can be divided into the stages of *inter-block* and *intra-block* pre-processing. Campos and Reis[159] have provided a systematic description of MB data pre-processing dividing it into three major steps:

- application of the standard chemometric pre-treatment operations to correct artefacts and uninteresting variations such as noise, multiplicative effects, scaling, baseline drift, peak shift and variations related to external factors (see Appendix A);
- equalization of the contribution of all variables *within* each block using classical methods such as mean-centering and unit variance scaling (see Appendix A);
- equalization of the *inter*-block systemic effects such as the differences in the scales, number of variables and the pseudo rank of different blocks. This level of pre-processing is necessary for those multi-block analysis methods that tend to favor the blocks with larger variations, leading to model bias. However, with proper scaling or weighting of blocks, it has been proven that model interpretation and predictive accuracy can be increased[160]. These inter-block pre-processing approaches are mainly scaling, such as *block scaling*, *block variance scaling* and *block rank scaling* (see Appendix A).

Although multi-block pre-processing is important, not all its levels are required in all multi-block data analysis approaches. *Sequential* multi-block approaches, i.e., methods that handle the multi-block data one block at a time involving orthogonalization steps, are, for example, less sensitive to the relative scaling of the blocks and can also deal with the differences in the ranks of multi-block data.

To deal with the challenges of visualizing multi-block multivariate data and using them for predictive purposes, several extensions of one-block chemometric methods as well as new specific multi-block approaches have emerged in recent decades. A detailed description of each of them is out of the aim of this chapter; only some general considerations will be provided.

Considering the main limitation of LDA approach, most of the methods for multi-block predictive analysis in chemometrics have been extensions of standard PLS (both for regression and classification purposes) to the multi-block scenario. On this regard, two were the firsts applications of PLS with multiple X blocks, a non-predictive method,

---

[159] Campos, M.P.; Reis, M.S. *Chemometr. Intell. Syst.* **2020**, *199*, 103959.
[160] Campos, M.P.; Sousa, R.; Pereira, A.C.; Reis, M.S. *Talanta*, **2017**, *171*,132-142.

called PLS path modeling[161], and a predictive multi-block PLS version[162], proposed with two different algorithms (averaging MB-PLS and stepwise). A refinement of these PLS methods, called Hierarchical two-block predictive PLS or PLS-2H, was introduced by Wold and collaborators[163], and other variations[164] of this H-PLS method have been proposed in literature since them. As well, others PLS algorithms for multiple blocks[165], based on the one originally presented by Wold and collaborators[166], were introduced. In particular, one of these algorithms, Sequential and Orthogonalized-PLS[167] (SO-PLS), has been used in this thesis.

A generic scheme of the SO-PLS algorithm, insertable among mid-level data fusion methods, is presented in Figure 3.9. The SO-PLS approach involves a series of standard PLS regression and matrix orthogonalization operations to extract sequentially the complementary information from different data-blocks. The term *sequential* means that blocks of data are incorporated, allowing to assess their incremental contribution. Sequential and Orthogonalized-Partial Least Squares-Linear-Discriminant Analysis (SO-PLS-LDA) is a multi-block classification method that derives from the association of SO-PLS and LDA. SO-PLS will be described in this paragraph in order to have an in-depth overview of the method in question.

---

[161] Gerlach, R.W.; Kowalski, B.R.; Wold, H.O.A. *Anal. Chim. Acta.* **1979**, *112*, 417-421.
[162] Frank, I.E.; Feikema, J.; Constantine, N.; Kowalski, B.R. *J. Chem. Info. Comput. Sci*. **1984**, *24*, 20-24.
[163] Wold, S.; Hellberg, S.; Lundstedt, T.; Sjostrom, M.; Wold, H. *Proc. Symp. on PLS Model Building: Theory and Application*. Frankfurt am Main, **1987**.
[164] Wold, S.; Kettaneh, N.; Tjessem, K. *J. Chemometr*. **1996**, *10*, 463-482.
[165] a) Qin, S.J.; Valle, S.; Piovoso, M.J. *J. Chemometr*. **2001**, *15*, 715-742; b) Lofstedt, T.; Trygg, J. *J. Chemometr*. **2011**, *25*, 441-455; c). Liland, K.H.; Næs, T.; Indahl, U.G. J. Chemometr. **2016**, *30*, 651-662.
[166] Wold, S.; Martens, H.; Wold, H. In Wold, S. (Ed.) *MULDAST Proc. Tech. Rep.* Research Group for Chemometrics, Umeå University, **1984**.
[167] a) Næs, T.; Tomic, O.; Mevik, B.-H.; Martens, H. *J. Chemom*. **2011**, *25*, 28-40; b) Næs, T.; Tomic, O.; Afseth, N.K.; Segtnan, V.; Måge, I. *Chemom. Intell. Lab. Syst.* **2013**, *124*, 32-42.

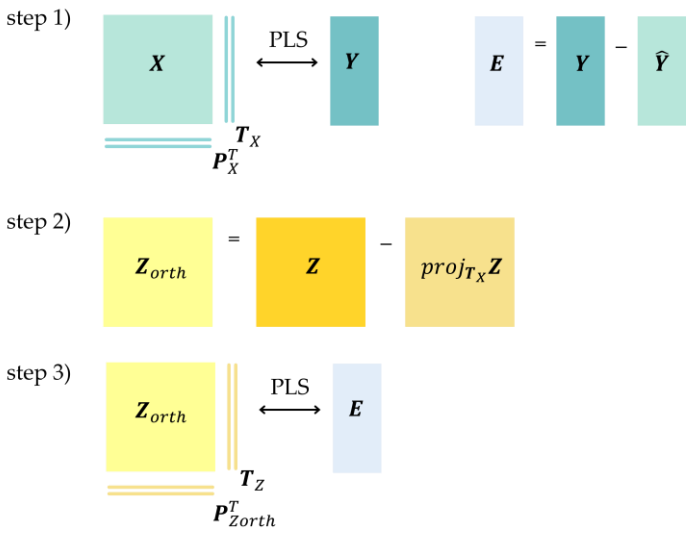


Figure 3.9. Simplified scheme of the orthogonalization step of the SO-PLS regression method.

For a SO-PLS model with two predictor blocks the model can be represented by eq. 3.39[168]:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{ZC} + \boldsymbol{E} \qquad \text{(eq. 3.39)}$$

where $\boldsymbol{Y}$ (NxK) is the response matrix, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the two data-blocks of dimensions (N×M_1) and (N×M_2), respectively, $\boldsymbol{B}$ (M_1×K) and $\boldsymbol{C}$ (M_2xK) are their regression coefficients, and $\boldsymbol{E}$ the residuals (N×K). This decomposition is achieved by means of four steps:

step 1) A PLS regression model is calculated between the first block $\boldsymbol{X}_1$ and the response matrix $\boldsymbol{Y}$, yielding $\boldsymbol{X}$-scores ($\boldsymbol{T}_X$), $\boldsymbol{X}$-loadings ($\boldsymbol{P}_X$), and $\boldsymbol{Y}$-loadings ($\boldsymbol{Q}_X$). The $\boldsymbol{Y}$-residuals matrix $\boldsymbol{E}$ is then calculated as:

$$\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{T}_X \boldsymbol{Q}_X^T \qquad \text{(eq. 3.40)}$$

step 2) The matrix $\boldsymbol{Z}$ is decomposed in $\boldsymbol{Z}_{orth}$ orthogonalized with respect to $\boldsymbol{T}_X$ and $\boldsymbol{Z}_{proj}$ projected on $\boldsymbol{T}_X$, i.e., $\boldsymbol{Z}_{orth}$ is obtained by deflating $\boldsymbol{Z}$ with respect to $\boldsymbol{T}_X$:

$$\boldsymbol{Z} = \boldsymbol{Z}_{orth} + \boldsymbol{Z}_{proj} \qquad \text{(eq.3.41)}$$

It follows that

---

[168] Biancolillo, A.; Næs, T. In Cocchi, M. (Ed.) *Data Fusion Methodology and Applications* (pp. 157-177), Part of *Data Handling in Science and Technology* Book Series, Volume 31. Elsevier, **2019**.

$$\boldsymbol{Z}_{orth} = \boldsymbol{Z} - \boldsymbol{Z}_{proj} = \boldsymbol{Z} - \boldsymbol{T}_X(\boldsymbol{T}_X^T\boldsymbol{T}_X)^{-1}\boldsymbol{T}_X^T\boldsymbol{Z} \qquad \text{(eq. 3.42)}$$

In practice, the orthogonalization subtracts $\boldsymbol{Z}_{proj}$ from $\boldsymbol{Z}$ by eliminating the part that lies in the column space of $\boldsymbol{T}_X$. This is the core of the method because it allows removing redundancies between the two predictor blocks.

step 3) Residuals from the first PLS regression (eq. 3.40) are then fitted to $\boldsymbol{Z}_{orth}$ by standard PLS. In this step the $\boldsymbol{Z}_{orth}$-scores ($\boldsymbol{T}_{Zorth}$), the $\boldsymbol{Z}_{orth}$-loadings ($\boldsymbol{P}_{Zorth}$), and the $\boldsymbol{Y}$-loadings $\boldsymbol{Q}_{Zorth}$ are calculated.

step 4) Owing to the orthogonality between $\boldsymbol{T}_X$ and $\boldsymbol{T}_{Zorth}$, $\boldsymbol{Y}$ can be predicted by summing the predictions from steps 1 and 3, as in eq. 3.43:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{T}_X\boldsymbol{Q}_X^T + \boldsymbol{T}_{Zorth}\boldsymbol{Q}_{Zorth}^T \qquad \text{(eq. 3.43)}$$

If more than two predictor blocks are tested, the algorithm follows by adding the further blocks taking care of orthogonalizing them with respect to all the other predictors already modelled. In a classification framework (SO-PLS-LDA), the $\boldsymbol{Y}$ real values, which were computed or predicted by the model, can be used to perform the classification basing on the probabilistic approach described before, including LDA. As well, LDA can be computed on the concatenated scores $\boldsymbol{T} = [\boldsymbol{T}_X\boldsymbol{T}_{Zorth}]$ of dimension (NxLVs$_{tot}$), where LVs$_{tot}$ is sum of the number of LVs of the first and second PLS. Among the different approaches for choosing the optimal complexity in SO-PLS, the *global approach*[167a] was used in this thesis. It is based on finding the set of components in all blocks that optimize the predictions. The graphical tool Måge plot has been developed to ease the choice of the optimal number of components when the global approach is used. On the horizontal axis the total number of components is shown and, in the plot itself, the combination of components corresponding to the different predictions are given. The best predictions for the sum of the two (or more if more blocks are used) components are indicated by the lower line. The inspection of global and local minima in the plot will suggest which combination of components leads to the best solution.

The major advantages of SO-PLS are linked to the orthogonalization, which removes redundant information, and to its sequential nature, which allows the interpretation of the incremental contributions provided by each data-block, sharing some resemblance with ANOVA.

SO-CovSel[169] is another sequential multi-block approach, based on the same algorithm of SO-PLS, in which the feature reduction operated by PLS is performed by a variable selection approach called Covariance Selection[170] (CovSel). Obviously, this leads to further slight differences in their algorithms. The first regression is calculated by

---

[169] Biancolillo, A.; Marini, F.; Roger, J.-M. *J. Chemom*. **2020**, *34(2)*, e3120.
[170] Roger, J.-M.; Palagos, B.; Bertrand, D.; Fernandez-Ahumada, E. *Chemom. Intell. Lab. Syst*. **2011**, *106(2)*, 216-223.

ordinary least square regression between $\boldsymbol{X}_{sel}$, the reduced $\boldsymbol{X}$-block obtained by the Covariance Selection, and $\boldsymbol{Y}$; in the same way, the selection is operated also on $\boldsymbol{Z}_{orth}$, providing $\boldsymbol{Z}_{orth,sel}$.

SO-CovSel, as well as SO-PLS, can be coupled with LDA to be applied in solving classification problems.

As mentioned above, pre-processing selection in Chemometrics is a challenging task. Powerful support comes from the emerging ensemble pre-processing fusion approaches, especially the SO-PLS inspired method such as Sequential pre-processing through ORThogonalization[171] (SPORT). In SPORT, several pre-treatments of the same data-block are associated in an SO-PLS model. In this scenario, considering two predictor blocks means dealing with only one predictor block $\boldsymbol{X}$ pre-treated by two different pre-processing approaches, resulting in the two pre-treated matrices, $\boldsymbol{X}_{pre-proc1}$ and $\boldsymbol{X}_{pre-proc2}$. Apart from this, the algorithm can be summarized by the same steps already described for SO-PLS.

[171] Roger, J.-M.; Biancolillo, A.; Marini, F. *Chemometr. Intell. Lab. Syst*. **2020**, *199*, 103975.

Abbreviations

$\overline{A}$

ANOVA, ANalysis Of VAriance

ASCA, ANOVA-Simultaneous Component Analysis

$\overline{C}$

CovSel, Covariance Selection

$\overline{D}$

DF, Data Fusion

DOE, Design of Experiments

$\overline{E}$

EDA, Exploratory Data Analysis

$\overline{H}$

H-PCA, Hierarchical Principal Component Analysis

$\overline{L}$

LDA, Linear Discriminant Analysis

$\overline{M}$

MANOVA, Multivariate Analysis of Variance

MB, Multi-Block

MB-PLS, Multi-Block Partial Least-Squares

MR, Multivariate Regression

$\overline{N}$

NIPALS, Nonlinear Iterative Partial Least Squares

$\overline{P}$

PCs, Principal Components

PCA, Principal Component Analysis

PLS, Partial Least Squares

PLS-DA, Partial Least Squares-Discriminant Analysis

$\overline{S}$

SC, Simultaneous Component

SCA, Simultaneous Component Analysis

SO-CovSel, Sequential and Orthogonalized-Covariance Selection

SO-CovSel-LDA, Sequential and Orthogonalized-Covariance Selection-Linear Discriminant Analysis

SO-PLS, Sequential and Orthogonalized-Partial Least Squares

SO-PLS-LDA, Sequential and Orthogonalized-Partial Least Squares-Linear Discriminant Analysis

SPORT, Sequential Pre-processing through ORThogonalization

SVD, Single Value Decomposition

$\overline{V}$

VIP, Variable Importance in Projection

Section 1

Multi-methodological characterization coupled with Chemometrics in the discrimination of typical Pecorino cheeses from Central Italy

S1.1 Introduction

Cheese is the generic name for a group of fermented milk-based food products; cheesemaking, indeed, refers to the production of cheese by using bacterial culture, enzymes and stabilizers to condense the milk proteins and fats and to preserve the cheese. Cheese has been produced since the earliest civilizations, 8000 years ago during the *Agricultural Revolution*. Cheesemaking was a farm-based operation until the mid-nineteenth century; nowadays the principal cheese varieties are manufactured in mechanized factories by highly developed technology, despite several cheeses are still produced at an artisanal level. The current production of cheese, which is the principal growth product within the dairy sector, is about $19*10^6$ tonnes per annum, predominantly in Europe, North and South America and Oceania, using approximately 35% of milk produced in the world. Europe has a predominant role, with a production of ~$11*10^6$ tonnes per annum[172]. The general cheesemaking process includes several steps: milk preparation, coagulation, cutting and cooking of the curd, its extraction (drainage of the whey), shaping of the drained curd, salting and ripening. These steps, schematized in Figure S1, outline the general principles of turning milk into cheese, common to most cheese varieties. Nevertheless, and despite of a basically similar raw material (milk from a few animal species), more than 1000 varieties of cheese[173] are nowadays available in commerce.

Figure S1. Main steps of the general cheesemaking process.

---

[172] Fox, P.F.; Guinee, T.P.; Cogan, T.M.; McSweeney, P.L.H. In Fox, P.F.; Guinee, T.P.; Cogan, T.M. and McSweeney, P.L.H. (Eds.) *Fundamentals of Cheese Science* (pp. 1-10), Springer, Boston, MA, **2017**.
[173] Sandine, W.E.; Elliker, P.R. *J. Agric. Food Chem*. **1970**, *18*, 557-566.

The *coagulation* of the main proteins in milk, i.e., caseins, is one of the most important process. Some proteolytic enzymes can modify the milk protein system, causing it to coagulate forming a gel in which the milk fat and aqueous phases are entrapped. Enzymes capable of promoting this transformation, referred to as *rennet*, are widespread and can be found in bacteria, molds, plants and slaughtered young animal stomachs (abomasa). Most modern cheese varieties and ~75% of the total world production[174] are obtained by rennet-coagulation. In Mediterranean countries, where sheep and goat breeding are largely diffused, the use of lamb or kid paste is a common practice. The active milk-curdling enzymes in animal rennet are chymosin and a different number of other generic proteases, such as pepsin A, B, or C. The most important difference among available rennet pastes is that lamb or kid rennet contains lipolytic enzymes for the hydrolysis of milk fat triglycerides yielding free fatty acids, whereas calf rennet doesn't exhibit lipolytic activity[175]. Commercial calf and bovine rennet, the most commonly used enzymes, are indeed lipolytic active free. The lipolytic agents are the only ones coming naturally from milk (milk lipases) and from microflora activity in cheeses coagulated with calf rennet[176].

The resulting milk gel is broken, separating into curds and whey. During the storage of these rennet-coagulated curds, referred to as *ripening* (maturation), the enzymes in rennet continue acting. Thus, a great diversity of flavors and textures can develop during storage, due to synchronized and balanced consecutive and concomitant biochemical events. Another source of variation in the characteristics of cheese is surely the animal species from which the milk is obtained. The cow is by far the principal worldwide source of the milk in commerce (85%), while 11%, 2% and 1.5% are produced by water buffalo, sheep and goats. Although sheep and goats are minor dairying species, they are quite significant in Mediterranean and Balkan countries.

---

[174] McSweeney, P.L.H.; Ottogalli, G.; Fox, P.F. In McSweeney, P.L.H.; Fox, P.F.; Cotter, P.D. and Everett, D.W. (Eds.) *Cheese*, Fourth Edition (pp. 781-808), Academic Press, **2017**.
[175] Addis, M.; Piredda, G.; Pirisi, A. *Small Rumin. Res*. **2008**, *79*, 2-10.
[176] Collins, Y.F.; McSweeney, P.L.H.; Wilkinson, M.G. *Int. Dairy J*. **2003**, *13*, 841-866.

S1.2 Aim and Outline of the project

Ewes' milk is used in the manufacture of more than 100 traditional cheese varieties in Europe, many of them with Protected Designation of Origin status. This ancestral heritage is a cultural trait of Mediterranean countries, contributing to the socioeconomic sustainability[177].

Ewes' milk cheese is a relatively common foodstuff produced in different countries all over the world. In 2017, 687511 tons of sheep milk cheese have been manufactured in the world, of these, approximately 56% of the global production (386968 tons) had origin in Europe and, on this regard, Italy is for sure one of the main players, producing, on its own, more sheep milk cheese than the amount manufactured in Eastern or Western Europe[178], covering 23% of the production in Southern Europe. A graphical comparison is shown in Figure S2.

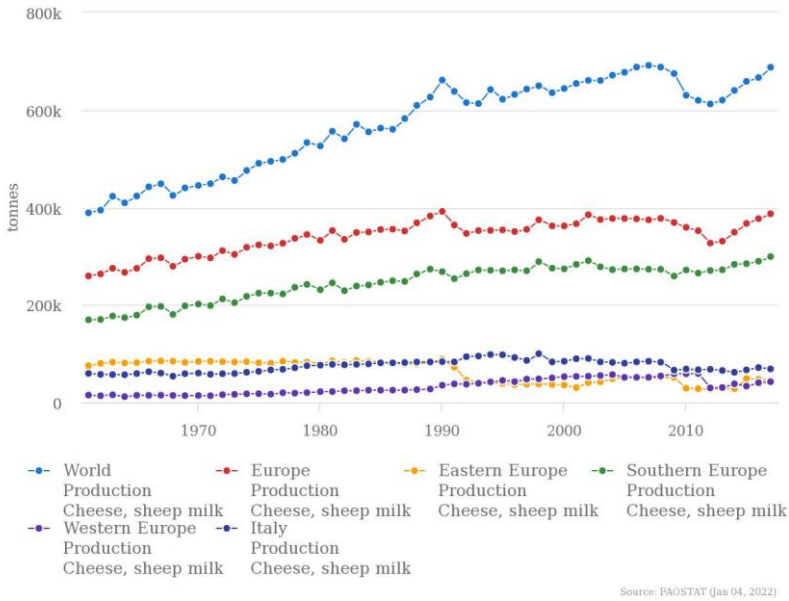


Figure S2. Comparison of the production of sheep milk cheeses in the world, in the different parts of Europe and in Italy. Time range is 1961-2017. More recent data are not available online. Reproduced from https://www.fao.org/faostat/en/#compare (last access January 2022).

---

[177] Medina, M.; Nuñez, M. In Fox, P.F.; McSweeney, P.L.H.; Cogan, T.M. and Guinee, T.P. (Eds.), *Cheese: Chemistry, Physics and Microbiology*, Third Edition, Volume 2: *Major Cheese Groups* (pp. 279-299), Amsterdam, The Netherlands: Elsevier, **2004**.

[178] FAO. FAOSTAT-Food and agriculture data from Food and Agriculture Organization of the United Nations. Retrieved from https://www.fao.org/faostat/en/ (last access January 2022).

The seasonal production of ewes' milk results in large variations in cheese production. Ewes' milk cheeses are produced in Southern Europe mainly between December and June, since milk production increases sharply in spring and decreases from July to November. For these reasons, procedures to circumvent variations in ewes' milk supply include freezing of unconcentrated milk[179] and freezing of fully-ripened cheeses[180].

Italian territory presents in fact an orographic structure suitable for ewes breeding, and, consequently, the production of sheep dairy commodities is a traditional and flourishing activity in several regions. For this reason, Italy yields a relatively high number of different sheep milk cheeses (called *Pecorino* cheeses from *pecora*, the Italian word for sheep) and several varieties have been awarded of the Protected Denomination of Origin (PDO), or other European and International quality marks. As mentioned, this product is manufactured in various Italian regions, especially in Southern and Central Italy and in the main islands, Sardinia and Sicily, where sheep's milk represents a common alternative to cow milk in the production of dairy products.

Generally, sheep's milk contains high levels of fat and protein (casein), which are the main cheesemaking constituents. The coagulum is firm, and NaCl diffusion is slow due to the low moisture content of these cheeses. After whey drainage and salting, cheese is ripened for a variable time, in which lipolytic, proteolytic and glycolytic processes generate a variety of chemical substances responsible for the appreciated taste and aroma. By modifying the cheesemaking technology, it is possible to obtain a range of cheeses. As concerning the maturation process, Pecorino cheeses are generally classified as *dolce* (soft-ripening), *maturo* (hard-ripening) or *da grattugia* (for grating). The final peculiar taste and flavor is closely related to the cheesemaking conditions, which often reflect the local or regional know-how (kind of starters and coagulants, curd heating temperature and ripening time) adopted in the production territory. This factor, together with the origin of milk, confers to Pecorino cheese a strong geographical identity. The tradition of making cheeses from sheep's milk is so rooted that these commodities follow productive procedures closely linked to ancient know-how of local shepherds, which makes these products highly peculiar, and, sometimes unique.

The present study takes the cue from the rich panorama of ewes' dairy products manufactured in Central Italy, and aims at the development of a multi-platform characterization of three typical Pecorino cheeses produced in Central Italy and certified with European (PDO) and International (Slow Food Presidia) quality schemes for tracing the origin and the cheesemaking procedure of these valuable commodities.

The bulls-eye has been focused on Pecorino di Farindola (PF), and two PDO specialties i.e., Pecorino Romano (PR) and Pecorino Sardo (PS) cheeses (see Figure S3).

---

[179] Anifantakis, E.; Kehagias, C.; Kotouza, E.; Kalantzopoulous, G. *Milchwis-senschaft* **1980**, *35*, 80-82.
[180] Tejada, L.; Sánchez, E.; Gómez, R.; Vioque, M.; Fernández-Salguero, J. *J. Food Sci*. **2002**, *67*, 126-129.

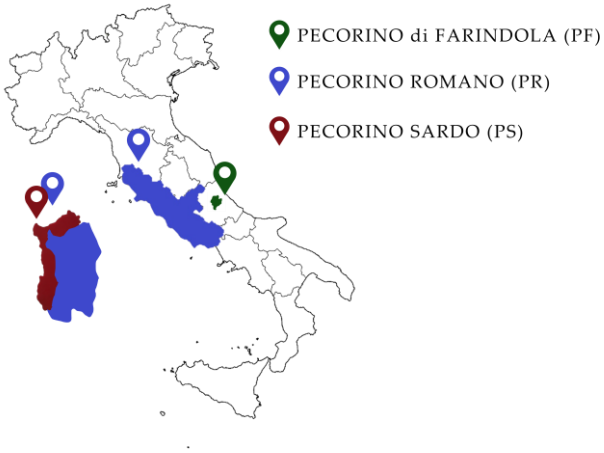


Figure S3. Geographical origin of Pecorino di Farindola (PF), Pecorino Romano (PR), and Pecorino Sardo (PS). The same colors-code and abbreviations will be used in all the chapters of Section 1.

PF, a Pecorino cheese produced in the Central Appenines (Abruzzo) near the homonymous town in the National Park of Gran Sasso and Monti della Laga, is the only one in the world coagulated with pig rennet (instead of the more common and stable calf or lamb ones) following a recipe which dates back to the Romans[181]. Due to its uniqueness[182], is a specialty protected by the Slow Food Presidium for the safeguard of typical food products.

PS is manufactured in Sardinia, whereas PR is mainly produced from milk coming from Sardinia (more than 95% in 2017) and Lazio (less than 5%)[183]. Despite the common origin of the raw materials, PS and PR have been selected because of their peculiar organoleptic characteristics that reflect the cheesemaking processes. PR, accounting for 80% of the total production of PDO ewe milk cheeses[184] and one of the most exported Italian cheeses in the world, is a semi-cooked hard cheese obtained by coagulation of whole ovine milk with lamb paste rennet coming from locally reared animals[185]. Going more deeply in the cheesemaking process, PR is inoculated with a natural starter culture (*scotta-fermento* or

---

[181] Slow Food Foundation for Biodiversity. Italian Slow Food Presidia. Retrieved from https://www.fondazioneslowfood.com/en/slow-food-presidia/farindola-pecorino/ (last access January 2022).

[182] a) Suzzi, G.; Sacchetti, G.; Patrignani, F.; Corsetti, A.; Tofalo, R.; Schirone, M.; Fasoli, G.; Gardini, F.; Perpetuini, G.; Lanciotti, R. *J. Sci. Food Agric.* **2015**, *95*, 2252-2263; b) Tofalo, R.; Schirone, M.; Fasoli, G.; Perpetuini, G.; Patrignani, F.; Manetta, A.C.; Lanciotti, R.; Corsetti, A.; Martino, G.; Suzzi, G. *Food Chem.* **2015**, *175*, 121-127.

[183] Alcuni dati sul Pecorino romano-Il Latte (italian source). Retrieved from http://www.lattenews.it/alcuni-dati-sul-pecorino-romano/ (last access January 2022).

[184] Pirisi, A.; Comunian, R.; Urgeghe, P.P.; Scintu, M.F. *Small Rum. Res.* **2011**, *101*, 102-112.

[185] Idda, I.; Spano, N.; Addis, M.; Galistu, G.; Ibba, I.; Nurchi, V.M.; Pilo, M.I.; Scintu, M.F.; Piredda, G.; Sanna, G. *J. Food Compos. Anal.* **2018**, *74*, 89-94.

scotta-innesto) obtained from the residual whey from Ricotta cheese manufacture. Coagulation with lamb rennet paste takes place at 37-39°C in 14-16 min. The coagulum is cut, left to settle for 2-3 min and cooked at 45-46°C. After 30 min, the curds are molded, pressed, and drilled to facilitate whey drainage. Cheeses are dry-salted periodically for 4-5 times over a period of about 30-60 days and ripened at 10-14°C for ripening time ranges from five to eight months.

PS is also a semi-cooked hard cheese, but it is inoculated with a commercial or natural whey starter culture and milk coagulation is promoted by addition of calf rennet. The curds are heated to 41-42°C and held for approximately 10 min. Cheeses are brine-salted for 48 h. It can be manufactured in two typologies, sweet (*dolce*) and mature (*maturo*), depending on the size of curd grains, on the size and weight of the wheels and their ripening status[186]. PS *dolce* is ripened for 20-60 days, weighs 1.0-2.3 kg and has aromatic and slightly acid flavor; the *mature* type, ripened for at least 4 months and weighing 1.7-4 kg, acquires with time a firm texture and a strong and piquant flavor. Specifications of PDO cheeses, generally, do not include rules about the use and the preparation of the rennet. Only for Pecorino Romano cheese is required that stomachs used in lamb rennet paste production come from animals raised in the PDO zone[187].

Because of their high commercial value, these three specialties may be subjected to frauds, for instance through the alteration of milk origin and cheesemaking process. Since cheese production is a relatively long and complex process, requiring controlled conditions, cheese adulteration cannot be easily mimed in laboratory. To overcome such limitation, the possibility of discriminating dairy products coming from different but well documented origins and technological procedures represent an efficient way to check the robustness of a given analytical method and the suitability of the selected markers to detect deliberate and fraudulent alterations of the milk origin and cheesemaking process. The three investigated Pecorino cheeses, apart from their good reputation and appreciation by consumers, seem to be particularly suitable for this purpose, because they are all certified specialties produced in Central Italy, namely under not very dissimilar pedologic and climatic conditions. In addition, the production territories partially overlap in the case of PS and PR cheeses, which permits to evaluate whether different cheesemaking conditions applied to milk and other raw materials coming from the same geographical area can affect the chemical composition of Pecorino cheeses.

Despite being a niche product, the peculiar attributes and high quality of Pecorino cheese have promoted different studies in the literature. As expected, some research works focused on its bacterial load, in particular, according to the thermal process of the starting milk (thermized or raw)[188], the kind of staring culture[186] and the type of rennet[182b].

---

[186] Mangia, N.P.; Murgia, M.A.; Garau, G.; Fancello, F.; Deiana, P. *Int. J. Dairy Technol*. **2013**, *66*, 543-551.
[187] Scintu, M.F.; Piredda, G. *Small Rumin. Res.* **2007**, *68*, 221-231.
[188] Randazzo, C.L.; Vaughan, E.E.; Caggia, C. *Int. J. Food Microbiol*. **2006**, *109*, 1-8.

Furthermore, much attention has been paid to the influence of the sheep diet[189], the time of production during the year[190], and cheesemaking and ripening procedures[191] on the organoleptic and physico-chemical properties of Pecorino cheeses. In this context, different analytical techniques were applied to classify Pecorino cheeses according to the ripening time[192].

S1.2.1 Why the volatile profile?

Cheese flavor results from the combination of hundreds of different compounds present in specific ratios and concentrations produced by three principal pathways during the ripening time. The involved metabolisms are in particular glycolysis, lipolysis, which lead to the liberation of free fatty acids, and proteolysis concerning the casein matrix degradation to peptides and free amino acids[193]. These substances directly affect the aroma, but they are also catabolized to give other flavored products. Lipolysis and proteolysis are catalyzed by a pattern of enzymes (lipases and proteinases and/or peptidases), which take origin from the milk, the rennet past used as coagulant and the starter and non-starter microbial species and are therefore characteristic of the cheese type. Gas-chromatography (GC) coupled with static or dynamic headspace solid-phase microextraction (HS-SPME) is the most suitable technique for the analysis of flavor compounds extracted directly from milk[194] or from dairy products coming from different countries[195]. SPME is known to be a sensitive, simple and solvent-free (no artifacts caused by the use of solvents) pre-analytical technique for the qualitative and semi-quantitative analysis of volatile compounds also in complex matrices, and this method has been widely adopted in food analysis since its first description[89]. However, a quantitative analysis by means of HS-SPME requires a rigorous standardization involving a set of different internal standards and the determination of factors addressing the different volatility of components. The HS-SPME profiling of cheese aroma was regarded as a promising tool for tracing the site of production, that may be different from the site of origin of the raw materials[196].

---

[189] Moran, L.; Aldezabal, A.; Aldai, N.; Barron, L.J.R. *Food Res. Int*. **2019**, *126*, 108669.
[190] Addis, M.; Fiori, M.; Riu, G.; Pes, M.; Salvatore, E.; Pirisi, A. *Small Rumin. Res*. **2015**, *126*, 73-79.
[191] a) Aldalur, A.; Ong, L.; Bustamante, M.Á.; Gras, S.L.; Barron, L.J.R. *Food Bioprod. Process*. **2019**, *116*, 160-169; b) Tekin, A.; Güler, Z. *Food Chem*. **2019**, *286*, 160-169.
[192] a) Lerma-García, M.J.; Gori, A.; Cerretani, L.; Simó-Alfonso, E.F.; Caboni, M.F. *J. Dairy Sci*. **2010**, *93*, 4490-4496; b) Cevoli, C.; Cerretani, L.; Gori, A.; Caboni, M.F.; Gallina Toschi, T.; Fabbri, A. *Food Chem*. **2011**, *129*, 1315-1319.
[193] McSweeney, P.L.H.; Sousa, M.J. *Lait*. **2000**, *80*, 293-324.
[194] a) Mounchili, A.; Wichtel, J.J.; Bosset, J.O.; Dohoo, I.R.; Imhof, M.; Altieri, D.; Mallia, S.; Stryhn, H. *Int. Dairy J*. **2005**, *15*, 1203-1215; b) Collomb, M.; Bütikofer, U.; Sieber, R.; Jeangros, B.; Bosset, J.O. *Int. Dairy J*. **2002**, *12*, 649-659.
[195] Pillonel, L.; Ampuero, S.; Tabacchi, R; Bosset, J.O. *Eur. Food Res. Technol*. **2003**, *216*, 179-183.
[196] Karoui, R.; De Baerdemaeker, J. *Food Chem*. **2007**, *102*, 621-640.

To attempt a discrimination of the above three typical Pecorino cheeses, representative samples belonging to each type were analyzed by GC with mass spectrometry (MS) detection combined with HS-SPME, and eventually the aroma profile was handled by discriminant analysis to classify the samples according to their geographical origin and cheesemaking process. This approach will be described in Chapter 4.

S1.2.2 Why the multi-elemental composition?

Compared to the profiling of molecular species or spectroscopic fingerprint methods, the determination of mineral and trace elements offers several advantages in food authentication and traceability[197]: i) the multi-elemental composition closely reflects the composition of the soil or the environment in which foodstuffs are produced, ii) it is stable and not substantially altered in the sample preparation, iii) a relatively simple digestion process is required as pre-analytical step, and iv) several elements can be simultaneously determined by atomic spectroscopy or spectrometry on plasma atomizers. In view of the above advantages, profiles of mineral and trace elements have been successfully utilized for the geographical discrimination of a wide variety of foodstuffs[198], but have been rarely used as authenticity or geographical markers of milk and dairy products[199]. Multi-elemental composition, in particular, was considered, alone or in combination with other instrumental responses, to discriminate milk produced in Malaysia[200] or in Southern Italy[201] from milk coming from other Countries. The most relevant investigation concerning cheeses, based on the combination of trace elements and isotopic ratios of biogenic elements (H, C, N and S), was aimed at differentiating PDO *Parmigiano Reggiano* from European and extra-European imitators, and predicting the origin of seven types of European hard cheeses[202]. Combination of isotopic and elemental profiles were also used to attempt the authentication of *Mozzarella di Bufala*

---

[197] Rodríguez-Bermúdez, R.; López-Alonso, M.; Miranda, M.; Fouz, R.; Orjales, I.; Herrero-Latorre, C. *Food Chem*. **2018**, *240*, 686-693.

[198] a) D'Archivio, A.A.; Foschi, M.; Aloia, R.; Maggi, M.A.; Rossi, L.; Ruggieri, F. *Food Chem*. **2019**, *275*, 333-338; b) Ruggiero, L.; Fontanella, M.C.; Amalfitano, C.; Beone, G.M.; Adamo, P. *Food Chem*. **2021**, *362*, 130168; c) Foschi, M.; D'Archivio, A.A.; Rossi, L. *Food Control*. **2020**, *118*, 107438; d) D'Archivio, A.A.; Di Vacri, M.L.; Ferrante, M.; Maggi, M.A.; Nisi, S.; Ruggieri, F. *Food Anal. Methods*. **2019**, *12*, 2572-2581; e) Varrà, M.O.; Husáková, L.; Patočka, J.; Ghidini, S.; Zanardi, E. *Food Chem*. **2021**, *356*, 129687; f) Wu, F.; Zhao, H.; Sun, J.; Guo, J.; Wu, L.; Xue, X.; Cao, W. *Food Chem*. **2021**, *354*, 129568.

[199] Kamal, M.; Karoui, R. *Trends Food Sci. Technol*. **2015**, *46*, 27-48.

[200] Zain, S.M.; Behkami, S.; Bakirdere, S.; Koki, I.B. *Food Control*. **2016**, *66*, 306-314.

[201] Sacco, D.; Brescia, M.A.; Sgaramella, A.; Casiello, G.; Buccolieri, A.; Ogrinc, N.; Sacco, A. *Food Chem*. **2009**, *114*, 1559-1563.

[202] Camin, F.; Wehrens, R.; Bertoldi, D.; Bontempo, L.; Ziller, L.; Perini, M.; Nicolini, G.; Nocetti, M.; Larcher, R. *Anal. Chim. Acta*. **2012**, *711*, 54-59.

Campana, a mozzarella cheese obtained from whole buffalo milk and certified by PDO mark[203].

In this thesis, the mineral composition of the selected dairy specialties was investigated by Inductively Coupled Plasma-Optical Emission Spectrometry (ICP-OES) and, eventually, the Pecorino cheeses were traced by Partial Least Squares-Discriminant Analysis. This approach will be described in Chapter 5.

S1.2.3 Why the fatty acids profile?

The primary components found in milk lipids are triacylglycerides (TAG), which account for more than 95% of the total lipid content[204]. As well as TAG, which contain three fatty acid (FA) chains, there are di- (two FA chains) and mono-acylglycerides (one FA chain). When a fatty acid is not bound to glycerol, it is termed a Free Fatty Acid (FFA). FFAs are primarily formed in dairy products by the enzymatic breakdown of glycerides[205]. The accurate determination of free fatty acids (FFA) in dairy products is important for research, legislative, process development and quality control, and its most immediate application is in measuring milk quality[206]. Moreover, the FFA content of dairy products is an important indicator of lipolysis in ripened and fermented products and is a useful marker when establishing and defining process conditions[207]. Lipolysis, i.e., hydrolysis of lipids, can be very extensive in hard-varieties of cheeses, such as Pecorino is. It is due to i) the action of the indigenous lipase in cheese made from raw milk, ii) the action of microbial lipases, or iii) the action of the lipases contained in the rennet paste used in the cheesemaking[208], during ripening. The FFA pattern found in cheeses, therefore, reflects that present in milk and depends on the composition of triglycerides. Moreover, as will be highlighted in Chapter 4, free fatty acids are also important precursors of catabolic reactions, which produce volatile compounds and contribute to cheese flavor[176].

GC with different detectors have been exploited to determine the authenticity of food by obtaining a fingerprinting chromatography profile[209]; fatty acid profile obtained by GC,

---

[203] a) Bontempo, L.; Barbero, A.; Bertoldi, D.; Camin, F.; Larcher, R.; Perini, M.; Sepulcri, A.; Zicarelli, L.; Piasentier, E. *Food Chem*. **2019**, *285*, 316-323; b) Brescia, M.A.; Monfreda, M.; Buccolieri, A.; Carrino, C. *Food Chem*. **2005**, *89*, 139-147.
[204] Jensen, R.G. *J. Dairy Sci.* **2001**, *85*, 295-350.
[205] Deeth, H.C.; Fitz-Gerald, C.H. In Fox, P.F. (Ed.) *Advanced Dairy Chemistry.Volume 2: Lipids* (pp. 247-308), London: Chapman and Hall, **1995**.
[206] Mannion, D.T.; Furey, A.; Kilcawley, K.N. *Int. J. Dairy Technol.* **2016**, *69*, 1-12.
[207] Fox, P.F.; Wallace, J.M. *Adv. Appl. Microbiol*. **1997**, *45*, 17-85.
[208] Gobbetti, M.; Di Cagno, R. In McSweeney, P.L.H.; Fox, P.F.; Cotter, P.D. and Everett, D.W. (Eds.), *Cheese: Chemistry, Physics and Microbiology*, Fourth Edition (pp. 809-828), Amsterdam, The Netherlands: Elsevier, **2017**.
[209] Cajka, T. In Ferrer, I. and Thurman, E. (Eds.) *Gas Chromatography-Time-of-Flight mass spectrometry in food and environmental analysis* (pp. 271-302), Amsterdam, The Netherlands: Elsevier, **2013**.

for example, has been used to determine the geographical origin of milk[210]. Gaspardo and collaborators[211] found that unsaturated and long-chain fatty acids can be used as efficient markers for the discrimination of milk samples based on the country of origin, in agreement with Palladino and co-workers[212], who have demonstrated that the variation of fatty acid composition in milk can be related with the source of origin and breed. Moreover, it has been assessed that the triacylglycerol profile of milk was different according to the feeding system (fresh grass feeding, pasture grazing and organic/biodynamic farming)[213]. Along this line, the FA profile of forages is known to be influenced by several factors such as systems, seasons, and processing. Moreover, the diversity of forages depends on farms, due to the different types of plant materials used to make forages and their ratio and quality[214].

In this thesis, the fingerprinting GC profile was determined for the TAG and FFA fractions to investigate its potentiality in the discrimination of the above-mentioned Pecorino cheeses. This approach will be described in Chapter 6.

S1.2.4 Why infrared-spectroscopy *fingerprinting*?

Despite these previous investigations will highlight the possibility of using specific chemical markers for discriminating Pecorino cheeses according to their manufacture and/or origin, the collection of the compositional data requires preliminary extraction or mineralization process, and sophisticated and expensive instrumentations, which may be unsuitable for routine quality controls. On the other hand, application of vibrational spectroscopies permits overcoming the above limitations and its potentiality has been amply demonstrated in the authentication or discrimination of PDO cow-milk cheeses, such as *Parmigiano Reggiano*[215], *Asiago*[216], and *Gruyére* and *L'Evitaz* Swiss cheeses[217].

The aim of using spectroscopy is to test whether is possible to develop a green, rapid, and non-destructive tool for discriminating PF, PR and PS varieties. In particular, representative samples belonging to the three categories of interest were analyzed by non-destructive Attenuated Total Reflectance-Fourier Transformed Infrared Spectroscopy (ATR-FTIR) technique. Eventually, the classification of the samples into PF,

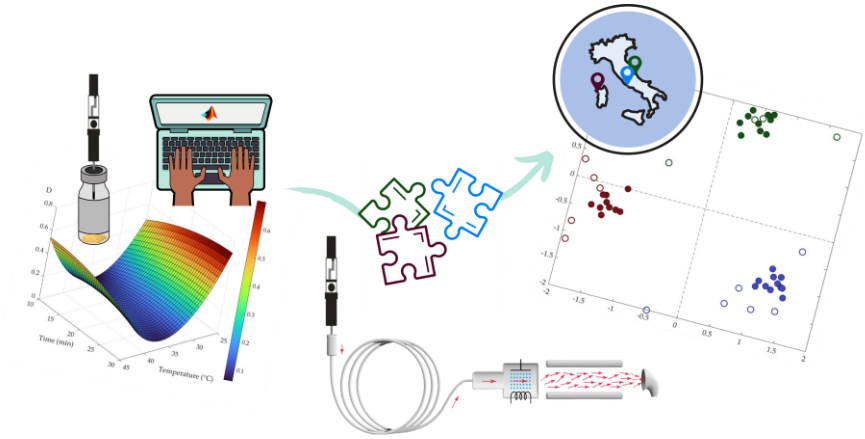
---

[210] Kamal, M.; Karoui, R. *Trends Food Sci. Technol.* **2015**, *46*, 27-48.
[211] Gaspardo, B.; Lavrenčič, A.; Levart, A.; Del Zotto, S.; Stefanon, B. *J. Dairy Sci.* **2010**, *93*, 3417-3426.
[212] Palladino, R.A.; Buckley, F.; Prendiville, R.; Murphy, J.J.; Callan, J.; Kenny, D.A. Journal of Dairy Science. **2010**, *93*, 2176-2184.
[213] Capuano, M.; Boerrigter-Eenling, R.; Elgersma, A.; van Ruth, S.M. *Eur. Food Res. Technol.* **2014**, *238*, 573-580.
[214] Lolli, V.; Renes, E.; Caligiani, A.; de la Fuente, M.A.; Gómez-Cortés, P. *J. Agric. Food Chem.* **2021**, *69*, 9654-9660.
[215] Gori, A.; Maggio, R.M.; Cerretani, L.; Nocetti, M.; Caboni, M.F. *Int. Dairy J.* **2012**, *23*, 115-120.
[216] Ottavian, M.; Facco, P.; Barolo, M.; Berzaghi, P.; Segato, S.; Novelli, E.; Balzan, Stefania. *J. Food Eng.* **2012**, *113(2)*, 289-298.
[217] Karoui, R.; Mazerolles, G.; Bosset, J.O.; de Baerdemaeker, J.; Dufour, E. *Food Chem.* **2007**, *105*, 847-854.

PR and PS classes was carried out by PLS-DA and an innovative ensemble learning approach, Sequential Pre-processing through ORThogonalization (SPORT) and LDA. This approach will be described in Chapter 7.

Chapter 4

HS-SPME/GC-MS volatile fraction determination and Chemometrics for the discrimination of typical Italian Pecorino cheeses[66]



4.1 Materials and Methods

4.1.1 Pecorino cheese samples and chemicals

Forty-seven (47) samples of Pecorino cheeses (14 PF, 16 PS and 17 PR) were acquired from local supermarkets or, when possible, directly from consortia (sampling involved 6 different producers of consortia for PF) to guarantee their authenticity, between September 2018 and April 2019. While PR, being mainly used for grating, is a hard cheese, both soft- and hard-ripening PF and PS cheese samples were collected to represent the products available in the market. In accordance with the respective regulations, the time of maturation ranges from a minimum of 40 days to over a year for PF cheese, from 20 to 60 days for PS cheese designated as *dolce,* is not less than two months for PS *maturo,* and ranges from 5 to 8 months for PR cheese. Anyway, no indication about how long it has been from the end of maturation to the purchasing is provided by the seller. This unchangeable condition allows testing the robustness of the further classification models. According to the cheesemaking protocols, milk and any other raw material originate from Sardinia in the case of PS, from Sardinia, Lazio and Province of Grosseto (Tuscany) for PR and from a restricted mountain area within the National Park of Gran Sasso and Monti della Laga located in the central Apennines

(Abruzzo) for PF (Figure S.3). Before the analysis, which was performed not later than 10 days from the purchase, the cheese samples were stored in a refrigerator at 4°C.

C_7-C_{40} Saturated Alkanes Standard dissolved in hexane from Sigma-Aldrich (Saint Louis, MO, USA) was used for the determination of the retention indices. Polydimethylsiloxane (PDMS), 100 μm thickness, carbowax/divinylbenzene (CW/DVB), 65 μm thickness, and 1 cm divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS), 50/30 μm thickness, fibers (Supelco, Bellafonte, PA, USA) were used in HS-SPME.

4.1.2 Headspace solid-phase micro-extraction (HS-SPME)

In order to ensure, as far as possible, the reproducibility of samples collection, the Pecorino cheese samples subjected to the analysis were extracted from the center of each cheese slice with an *in-house* designed polypropylene device. Aliquots of 3 mg (±0.01 mg accuracy) were placed in a 12-mL vial capped with a PTFE-silicon septum and allowed equilibrating in a water bath kept at constant temperature (±1°C accuracy) for 30 min. The SPME fiber, previously conditioned in the chromatograph injector at the temperature suggested by the manufacturer (250°C for PDMS, 220°C for CW/DVB and 270°C for DVB/CAR/PDMS fiber), was placed in the headspace of the sample and let to interact with the volatile compounds. The extraction efficiency of PDMS, CW/DVB and DVB/CAR/PDMS fibers were at first evaluated in the same conditions in order to select the most suitable sorbent for the analysis of volatile compounds in Pecorino. The analyses were conducted on the sample kept at 45°C and the fiber was exposed for 30 min. After sampling, desorption of the volatile compounds took place in the GC injection port for 5 min at 250°C.

4.1.3 Gas chromatography-mass spectrometry (GC-MS) analysis

GC-MS analysis of the extracted volatile compounds was performed on a Varian Saturn 2000 (Varian, Inc, Walnut Creek, CA, USA) GC–MS system composed by a Star GC 3400 CX gas chromatograph, equipped with a 1078 split/splitless injector and a SPME liner inside, connected to an ion-trap mass detector. A Varian FactorFour™ VF-WAX capillary column (30 m × 0.25 mm × 0.25μm film thickness) was used, and the carrier gas was Helium IP supplied-at a flow rate of 1.0 mL/min. The column temperature was held at 35°C for 5 min, then increased to 150°C at a rate of 5°C/min and to 250°C at 10°C/min and held for 17 min. The injector temperature was held at 250°C and all the injections were carried out in the splitless (100:1 split ratio) mode. The presence of carry-over effects was monitored after each injection by a blank analysis (injection of air). The chromatograms were acquired as total ion current (TIC) values with a time resolution of 0.96 sec. The retention indices of the Pecorino volatiles were determined from the

respective retention times and the retention times of n-alkanes (C_7-C_{40}), collected in the same chromatographic conditions.

4.1.4 Statistical Analysis

4.1.4.1 Optimization of HS-SPME extraction by design of the experiments (DOE) and Response Surface Methodology (RSM)

In order to evaluate the combined effect of the sample temperature (T) and the exposure time (t) to the sample headspace (i.e., the two factors which are expected to mostly affect the extraction of volatiles) on the HS-SPME efficiency, a three-level full-factorial design was applied to plan the experiments. Each variable was investigated at three levels, and two replicate measurements of the central point were found to be sufficiently to ensure a low leverage of the model. The mathematical model for the response was assumed as follows:

$$Y = b_0 + b_1x_1 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2 \qquad \text{(eq. 4.1)}$$

where $x_1 = T$ and $x_2 = t$.

Regardless of the experimental conditions, the chromatograms collected in the various DOE points presented five very intense peaks. To enhance information of the overall chromatogram, two responses, Y_1 and Y_2, were maximized:

Y_1: overall percentage area of the less intense signals;

Y_2: number of detectable signals.

The mathematical approach of desirability function[218] was applied to convert the two-response optimization problem into a single-response one. A partial desirability function (d_i) defining an index of acceptability, ranging between 0 and 1, was defined by fixing the lower and upper acceptable value for each of the two responses Y_i. A global desirability function (D) was then calculated as the geometric mean of the two di values and used as the only response. The response surface generated by regression of the observed D values in the points of the DOE against the factors was finally analyzed to find the optimal condition. The model computation was performed using the R-based chemometric software CAT[219].

---

[218] Candioti, L.V.; De Zan, M.M.; Cámara, M.S.; Goicoechea, H.C. *Talanta* **2014**, *124*, 123-138.

[219] Leardi, R.; Melzi, C.; Polotti, G. CAT (Chemometric Agile Tool). Retrieved from http://gruppochemiometria.it/index.php/software (last access January 2022).

4.1.4.2 Discrimination of Pecorino cheeses

The peak areas of the volatile compounds were obtained by analyzing the GC chromatograms with the Varian Workstation Toolbar (Varian Inc., Walnut Creek, CA, USA). The resulting data-tables were exported to the Matlab environment (R2020b; The Mathworks, Natick, MA), to be further processed. The problem of classifying the samples according to their geographical origin and cheesemaking protocol was coped with LDA and PLS-DA. LDA was performed using the MATLAB classification toolbox[220] released by Milano Chemometrics and QSAR Research Group; whereas the Matlab function for PLS-DA here employed can be freely downloaded from the RomeChemometrics website [221]. For further details about the theory of the mentioned Chemometric techniques, the reader can refer to Chapter 3.

4.2 Results and Discussion

4.2.1 Investigation by DOE-RSM of the SPME efficiency

As a first step of this study, the three different mentioned fibers were tested in the analysis of Pecorino cheese volatiles ~~in~~ under the same HS-SPME/GC–MS conditions, defined in Section 4.1.2.

Based on the qualitative comparison of the obtained chromatograms (Figure 4.1), the DVB/CAR/PDMS coating turns out, in line with the literature[222,192b], to be clearly the most appropriate for the adsorption of the compounds characterizing the cheese aroma.

---

[220] Ballabio, D.; Consonni, V. *Anal. Methods*. **2013**, *5*, 3790-3798.
[221] Algorithms|RomeChemometrics. Retrieved from https://www.chem.uniroma1.it/romechemometrics/-research/algorithms/ (last access January 2022).
[222] Randazzo, C.L.; Pitino, I.; De Luca, S.; Scifò, G.O.; Caggia, C. *Int. J. Food Microbiol*. **2008**, *122*, 269-278.

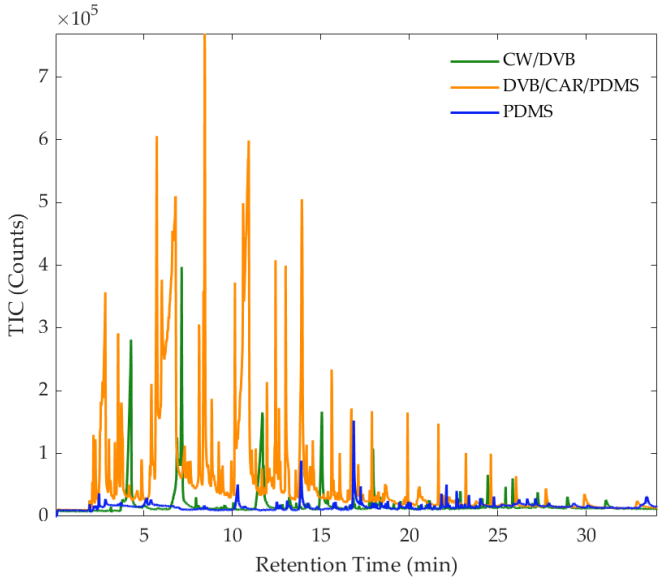


Figure 4.1. Comparison of GC chromatograms collected with three different SPME fibers from a PS cheese sample.

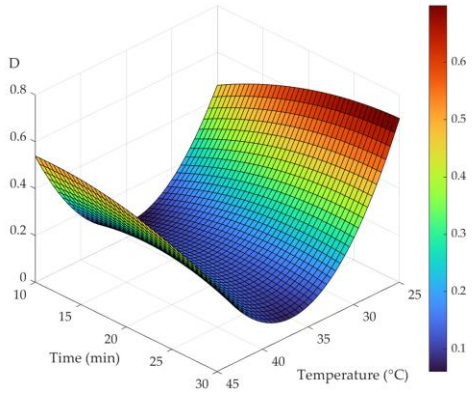
This fiber can adsorb compounds with a wide range of polarities thanks to the combination of the three different stationary phases. The porous carboxen layer in particular can capture the smaller and highly volatile compounds whereas the larger and less volatile substances can be extracted by the porous DVB sorbent. Moreover, inter-analyte displacement is very limited in DVB/CAR/PDMS fibers[223].

To maximize chemical information on Pecorino volatiles, the extraction performance of the DVB/CAR/PDMS sorbent was optimized by investigating the combined effect of sample temperature (T) and exposure time (t) of the fiber to the headspace. According to the selected design, each factor was varied within three levels: 25, 35, 45°C for T and 10, 20, 30 min for t. This choice for the levels is justified as follows. It was not possible to set T below room temperature and, on the other side, the sample will melt above 45°C. Time constraints are given by the need for a minimum time that allows the return of the GC from the fiber conditioning temperature (see Section 4.1.2) to the initial one before the analysis (see Section 4.1.3) and the intent of reducing, as far as possible, the overall analysis time. The experiments were carried out in a random order and single desirabilities were calculated in each point of the design assuming 10/50 and 50/75 as lower/upper acceptable values for the percentage residual area (Y_1) and the number of detectable peaks (Y_2), respectively. These quantities were combined in the global

[223] Risticevic, S.; Pawliszyn, J. *Anal. Chem.* **2013**, *85*, 8987-8995.

desirability. The full second-order model obtained by regression of the D values against the SPME factors is reported in Eq. 4.2, where the standard deviations of the model coefficients are given in brackets, while the surface plot for D and the matrix of experiments are reported in Figure 4.2.

$$D = (0.13 \pm 0.04) - (0.06 \pm 0.02)T + (0.02 \pm 0.02)t - (0.11 \pm 0.03)Tt + (0.44 \pm 0.04)T^2 - (0.05 \pm 0.04)t^2 \quad \text{(eq. 4.2)}$$



T (°C)	t (min)	d_1	d_2	D_{exp}	D_{calc}
-1	-1	0.72	0.36	0.51	0.46
-1	0	1.00	0.32	0.57	0.63
-1	1	0.51	1.00	0.71	0.70
0	-1	0.05	0	0	0.06
0	0	0.16	0.20	0.18	0.13
0	1	0.02	0.56	0.11	0.09
1	-1	0.30	1.00	0.55	0.54
1	0	0.75	0.36	0.52	0.50
1	1	0.31	0.36	0.33	0.34
0	0	0.09	0.16	0.12	0.13

Figure 4.2. Response surface describing global desirability (D) as a function of sample temperature (T) and fiber exposure time (t) in HS-SPME (on the left) and matrix of experiments with values of single desirabilities (d_1 and d_2) and experimental (D_{exp}) and calculated (D_{calc}) global desirability (on the right).

Based on Student's statistics, the T^2 term in model (eq. 4.2) is highly significant ($p < 0.001$) while a significant interaction (p<0.05) between the two factors also occurs, which does not allow simplifying the surface model.

The good descriptive and predictive abilities of the response surface are supported by the high value of the determination coefficient in calibration (R^2=0.94) and by the acceptable value of the determination coefficient in leave-one-out cross-validation (Q^2_{cv}=0.78). In addition, ANOVA test (for more details on ANOVA results, refer to Table 4.1) revealed that the surface model was highly significant (p=0.0032) while lack of fit was non-significant (p=0.4375). The maximum value of desirability inside the experimental domain was achieved in the (-1,1) point of the design, namely at T=25°C and t=30 min, indicating the lowest level is optimal for the factor T, whereas the highest one is required for the variable t. The shape of the convex surface suggests that a further decrease of temperature combined with a further increase of time can lead to a further increase of D. Nevertheless, to avoid an excessive increase of the analysis time and to overcome difficulty in the control of the sample temperature below the room temperature, the

SPME conditions in next experiments were set to T=25°C and t=30 min. The same conclusions came out by the application of Pareto front.

Table 4.1. ANOVA table (including all factors and their sum of squares, degree of freedom, mean square, F ratio and p values) for the experimental design; * indicate the level of significance (*p-value<0.05, **p-value<0.01, ***p-value<0.001).

ANOVA Table of the response model					
Source	**Sum of squares (SS)**	**Degree of freedom**	**Mean square (MS)**	**F ratio**	**p value**
Model	0.5191	5	0.1038	28.42	0.0032**
Error	0.0146	4	0.0037		
Total	0.5338	9			
Lack-of-fit test					
Source	**Sum of squares (SS)**	**Degree of freedom**	**Mean square (MS)**	**F ratio**	**p value**
Lack-of-fit	0.0128	3	0.0043	2.37	0.4375
Pure error	0.0018	1	0.0018		
Total error	0.0146	4			
Effect of parameters					
Source	**Sum of squares (SS)**	**Degree of freedom**	**Mean square (MS)**	**F ratio**	**p value**
Intercept	0.0462	1	0.0463	12.67	0.0236*
T	0.0253	1	0.0254	6.94	0.0579
t	0.0013	1	0.0014	0.37	0.5761
$T * t$	0.0441	1	0.0441	12.07	0.0255*
T^2	0.4473	1	0.4473	122.44	0.0004***
t^2	0.0063	1	0.0063	1.74	0.2580

4.2.2 HS-SPME/GC-MS profile of Pecorino cheese aroma

The chemical structure of the volatile components detected in the GC chromatogram was assigned by comparison of the observed mass spectra and retention indices with literature data or NIST14 library. The list of volatiles is provided in Table 4.2 together with a comparison of volatile profiles of the three Pecorino cheese varieties here investigate. The compounds determined in the Pecorino aroma acids, carbonyl compounds (in which ketones largely prevail on aldehydes), esters, alcohols and minor components such as terpenes, ethers and olefins.

Table 4.2. List of volatiles detected in the Pecorino cheese aroma with identification numbers (ID), observed retention times (RT), mean peak area (%) and related standard error in each of the three varieties, retention indices (RI), RI data from literature (RI (lit)) and mass spectral data. The first fragment ion in the list is the one with the higher *m/z* whereas the ion in bold is the one with the higher relative intensity. The remaining ions are listed in decreasing order of *m/z*.

		RT (min)	PF (n=14)	PS (n=16)	PR (n=17)	RI	RI (lit)	Characteristic Fragment Ions (m/z)
ID	**Acids**		**73 (±5)**	**71 (±6)**	**84 (±4)**			
26	Acetic acid	21.38	35±4	27±3	15±2	1440	1449	60, **43**
29	Propanoic acid	23.82	3.1±1.7	0.3±0.2	0.6±0.2	1537	1535	**74**, 57, 45, 30
30	2-methylpropanoic acid	24.55	1.3±0.3	0.75±0.15	0.18±0.09	1567	1570	88, **43**, 73, 55
31	Butanoic acid	26.02	28±2	38±5	62±3	1628	1625	**60**, 73, 55, 41
33	3-methylbutanoic acid	27.03	1.8±0.4	0.80±0.17	0.4±0.2	1670	1666	87, **60**, 43
34	Pentanoic acid	28.61	0.26±0.05	0.23±0.03	0.25±0.03	1742	1733	73, **60**, 55, 41
38	Hexanoic acid	30.64	3.6±0.9	3.7±0.7	5.3±1.4	1848	1846	87, **60**, 73, 55, 41
	Carbonyl compounds		**9 (±2)**	**17 (±6)**	**4 (±1)**			
1	Propan-2-one	3.36	0.9±0.5	0.16±0.06	0.22±0.11	810	819	58, **43**
4	Butan-2-one	4.62	0.39±0.14	n.d.[a]	0.14±0.07	893	907	72, **43**, 57
7	Pentan-2-one	6.64	1.7±0.3	0.45±0.15	0.24±0.12	965	981	86, **43**, 71, 58
8	Butane-2,3-dione (diacetyl)	6.78	0.09±0.04	0.7±0.3	0.11±0.06	969	970	86, **43**
13	Heptan-2-one	13.41	2.3±0.7	1.1±0.5	0.29±0.13	1165	1182	71, **43**, 58
18	3-hydroxybutan-2-one (acetoin)	16.66	1.4±0.4	14±6	3.0±0.9	1269	1284	88, **45**
22	Nonan-2-one	19.77	2.4±1.4	1.0±0.6	0.14±0.05	1379	1390	142, **43**, 127, 113, 99, 85, 71, 58
23	Nonanal	19.91	0.05±0.02	0.14±0.07	0.07±0.06	1384	1391	95, **41**, 82, 70, 57
28	Benzaldehyde	23.41	n.d.	0.06±0.03	0.10±0.07	1520	1520	**106**, 77, 51
	Esters		**3.0 (±0.6)**	**2.0 (±0.6)**	**1.4 (±0.3)**			
3	Ethyl acetate	4.39	0.27±0.08	0.10±0.03	n.d.	878	888	70, **43**, 61
10	Ethyl butanoate	8.54	1.1±0.3	0.45±0.17	0.23±0.09	1024	1035	101, **43**, 88, 71, 60
16	Butyl butanoate	14.66	n.d.	n.d.	0.10±0.07	1203	1220	89, **71**, 56, 41
17	Ethyl hexanoate	15.14	1.0±0.4	0.27±0.09	0.14±0.03	1219	1233	115, **88**, 99, 73, 60, 55, 43
19	2-methyl butyl isovalerate	17.00	0.059±0.013	0.15±0.08	0.11±0.07	1280	1299	85, **70**, 55 43
24	Ethyl heptanoate	20.08	0.36±0.23	0.53±0.19	0.35±0.08	1390	1375	113, **88**, 101, 73, 60, 55, 43
25	Ethyl octanoate	21.02	0.25±0.07	0.23±0.04	0.038±0.014	1426	1435	143, **88**, 127, 115, 101, 73, 57, 41
32	Ethyl decanoate	26.28	n.d.	0.7±0.5	0.4±0.3	1639	1638	155, **88**, 101, 73, 60, 55, 43

	Alcohols		**7.0 (±0.8)**	**4.0 (±1.2)**	**1.6 (±0.4)**			
5	Propan-2-ol	5.33	0.3±0.1	0.023±0.013	n.d.	920	927	59, **45**
6	Ethanol	5.50	0.7±0.2	0.45±0.16	0.08±0.02	926	932	45, **31**
9	Butan-2-ol	8.20	1.3±0.5	0.10±0.07	0.21±0.15	1015	1025	59, **45**
11	Pentan-2-ol	11.44	2.1±0.5	1.0±0.7	0.004±0.003	1107	1119	73, **45**, 55
12	Butan-1-ol	12.19	0.07±0.03	0.042±0.016	0.11±0.06	1129	1142	56, **41**, 31
15	3-methylbutan-1-ol	14.26	1.1±0.3	0.07±0.03	0.021±0.012	1191	1209	70, **55**, 41, 31
20	Heptan-2-ol	17.74	0.72±0.19	0.10±0.03	0.02±0.01	1304	1320	83, **45**, 70, 55
27	2-propylpentan-1-ol	22.48	0.18±0.05	1.8±0.9	0.9±0.3	1483	1485	83, **57**, 70, 41
36	2-phenylpropan-2-ol	28.98	n.d.	0.5±0.2	0.20±0.09	1760	1773	121, **43**, 118, 77
	Others		**8 (±2)**	**6 (±2)**	**9 (±3)**			
	Ethers		7±2	6±2	8±3			
21	Anisole	18.40	0.33±0.17	0.27±0.12	0.03±0.02	1329	1347	**108**, 93, 78, 65, 51, 39
35	Dicaprylyl ether	28.88	0.009±0.007	0.16±0.04	0.04±0.01	1755	1761	112, **71**, 84, 57, 43
37	Unknown#1 (ether)	30.60	6±2	5.1±1.5	8±3	1846	-	113, **69**, 41
	Terpenes		0.05±0.02	0.21±0.09	0.3±0.2			
14	Limonene	13.84	0.05±0.02	0.08±0.03	0.16±0.09	1178	1199	136, **68**, 121, 107, 93, 79, 53, 39
39	Geranyl acetone	30.85	n.d.	0.13±0.06	0.14±0.11	1859	1859	151, **43**, 136, 125, 107, 93, 69
	Olefins		0.6±0.2	0.4±0.2	0.23±0.11			
2	3-octene	4.26	0.6±0.2	0.4±0.2	0.23±0.11	869	850	112, **41**, 83, 70, 55

[a]: not detected

It must be reminded that the three Pecorino cheeses investigated in this work are produced using different kinds of coagulant, namely calf (PS), lamb (PR), and pig (FR) rennet, which is considered a relevant factor in the cheesemaking process and is expected to influence the evolution of cheese sensorial properties[224].

4.2.2.1 Acids

Free carboxylic acids (FFAs) are the major volatile components of all the three Pecorinos, especially PR in agreement with higher ripening. In fact, PR can be considered a prevalent lipolytic ripening sheep cheese, like others Italian hard cheeses including Fiore Sardo and Pecorino Siciliano, in which hydrolysis of triglycerides is promoted by the lipolytic enzymes of the lamb rennet[225]. The pattern of FFA found in cheeses obviously reflects that of the original milk and depends on the composition of triglycerides. In milk, the lateral position of the glycerol backbone in triglycerides (*sn*-1 or *sn*-3) is predominately occupied by short-chain linear fatty acids, especially butanoic and hexanoic acids, and unsaturated fatty acids, whereas the central (*sn*-2) position is preferentially bound to longer chain saturated fatty acids. Pregastric esterase contained in the rennet paste is highly specific for short chain acids esterified at the *sn*-3 position of glycerol, which explains the high content of butanoic acid and, to a lesser extent, of hexanoic acid in the aroma of the three Pecorino cheeses here investigated. Acetic acid is an important flavor compound that can originate from the action of lactic acid bacteria on lactose, the action of citrate-positive bacteria on citrate and the catabolism of amino acids. 2-methylpropanoic and 3-methylbutanoic acids were also found in this work. These branched chain acids were previously found in cheese aroma[226,182a] that originate from valine and leucine catabolism[227].

4.2.2.2 Carbonyl compounds

Among carbonyl compounds, ketones, especially alkan-2-ones, are the most represented. PF in particular presents higher contents of alkan-2-ones with odd number of carbon atoms (C_{n-1}), that are produced by β-oxidation of fatty acids (C_n) to β-ketoacids and subsequent decarboxylation, whereas acetoin that, together with diacetyl, is an important flavoring compound deriving from citrate metabolism prevails in PS and PR. Aldehydes can be formed by amino acid catabolism during cheese ripening involving

---

[224] Di Giacomo, F.; Casolani, N.; Del Signore, A. *Ital. J. Food Sci*. **2014**, *26*, 153-161.
[225] Urgeghe, P.P.; Piga, C.; Addis, M.; Di Salvo, R.; Piredda, G.; Scintu, M.F.; Wolf, I.V.; Sanna, G. *Food Anal. Methods*. **2012**, *5*, 723-730.
[226] Coda, R.; Brechany, E.; De Angelis, M.; De Candia, S.; Di Cagno, R.; Gobbetti, M. *J. Dairy Sci*. **2006**, *89*, 4126-4143.
[227] Yvon, M., Rijnen, L. *Int. Dairy J*. **2001**, *11*, 185-201.

Strecker degradation and as secondary products of autoxidation of unsaturated fatty acids[228]. Benzaldehyde is produced by β-oxidation of cinnamic acid or α-oxidation of phenylacetaldehyde[229]. However, aldehydes do not accumulate in cheeses because are rapidly converted into alcohols or into the corresponding acids[230], which justifies the low content of these compounds in the volatile fraction of the Pecorino cheeses here investigated.

4.2.2.3 Esters

Esters in cheese aroma originate from reaction between FFA and alcohols, ethyl alcohol in particular[231]. Ethyl butanoate and ethyl hexanoate, key odorant compounds having low perception threshold, prevail in PF headspace while ethyl heptanoate and ethyl decanoate, the latter being totally absent in PF, are the most abundant in PS and PR. Butyl butanoate was detected only in PR. The formation of 2-methyl butyl isovalerate, previously reported in Cheddar cheese, is justified by the presence of both 3-methylbutanoic acid and 2-methylbutan-1-ol, developed from leucine and isoleucine, respectively, through the catabolism of branched-chain FAAs[232].

4.2.2.4 Alcohols

Alkan-2-ols, originated by reduction of the corresponding alkan-2-ones, are the most abundant alcohols, especially in PR aroma. Moreover, we detected some primary alcohols such as butan-1-ol and 3-methylbutan-1-ol, deriving from catabolism of branched chain amino acids[233,207]. PS and PR aroma also contains 2-propylpentan-1-ol and 2-phenylpropan-2-ol. The latter compound is structurally similar to 2-phenylethanol, already identified in cheeses and produced in the phenylalanine catabolism[233a].

---

[228] Barbieri, G.; Bolzoni, L.; Careri, M.; Mangia, A.; Parolari, G.; Spagnoli, S.; Virgili, R. *J. Agric. Food Chem*. **1994**, *42*, 1170–1176.
[229] Sieber, R.; Bütikofer, U; & Bosset, J.O. *Int. Dairy J*. **1995**, *5*, 227–246.
[230] Ziino, M.; Condurso, C.; Romeo, V.; Giuffrida, D.; Verzera, A. *Int. Dairy J*. **2005**, *15*, 585–593.
[231] Cadwallader, K.R.; Singh, T.K. In McSweeney, P.L.H. and Fox, P.F. (Eds.), *Advanced Dairy Chemistry: Lactose, Water, Salts and Minor Constituents*, Volume 3 (pp. 631-690), Switzerland: Springer Nature, **2009**.
[232] Singh, T.K.; Drake, M.A.; Cadwallader, K.R. *Compr. Rev. Food Sci. Food Saf*. **2003**, *2*, 166–189.
[233] a) Bertuzzi, A.S.; McSweeney, P.L.H.; Rea, M.C.; Kilcawley, K.N. *Compr. Rev. Food Sci. Food Saf*. **2018**, *17*, 371–390.

4.2.2.5 Others

Among ethers, anisole, whose derivatives were previously found in cheese aroma[234], and dicaprylyl ether were detected. Limonene and geranyl acetone detected in this work are terpenes naturally present in milk before cheese manufacture due to their content in feeding[208,228]. Among hydrocarbons, that can be generated in lipid autoxidation during cheese ripening or derive directly from feeding stuff[222,224], 3-octene was the only one detected in this work.

4.2.3 PCA exploration of GC data

Forty-seven samples of Pecorino cheeses were analyzed by means of HS-SPME/GC–MS under the optimized conditions to attempt a classification based on the volatile compounds extracted by the DVB/CAR/PDMS fiber. PCA was performed on the peak areas of the 39 identified volatile compounds. The non-identified volatiles by contrast were not included in statistical analysis in order to guarantee a chemical interpretation of the classification models. A preliminary data exploration on both absolute and relative (%) peak areas revealed a heterogeneous distribution of the samples in the PC subspace regardless of cheese origin and manufacturing procedure. With the aim of increasing the inter-class differentiation of the Pecorino cheeses, a logarithmic transformation was finally applied to the non-null absolute peak areas of the volatiles. This transformation also reduced skewness of variable distribution.

---

[234] Frank, D.C.; Owen, C.M.; Patterson, J. *LWT-Food Sci. Technol*. **2004**, *37*, 139-154.

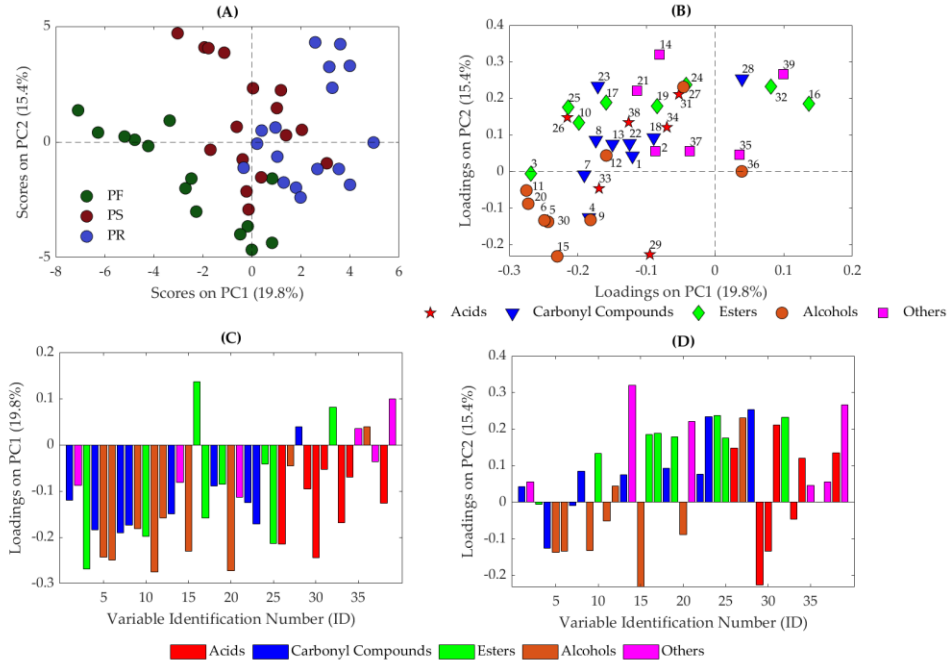


Figure 4.3. PCA Analysis. Projection of Pecorino samples (A) and chromatographic descriptors (B) onto the PC1-PC2 plane. Plot of loadings on PC1 (C) and on PC2 (D) as a function of the variables. The volatiles in (B, C, D) are coded according to the chemical family they belong to and identified by numbers (ID) reported in Table 4.2.

The projection of the Pecorino samples in the plane of the two first components (PC1 and PC2) extracted from the auto-scaled data previously subjected to logarithmic transformation (Figure 4.3(A)) shows a clear differentiation (mainly along PC1) of the PF samples and a less pronounced discrimination of the other two Pecorino types. The less evident differentiation of PS and PR samples apparently reflects the common origin of milk, which is exclusively (PS) or predominantly (PR) produced in Sardinia region. The distribution of the variables in the loadings plot (Figure 4.3(B)) can help to interpret the sample behavior. In particular, the evident separation of the PF samples from PS+PR classes along the bi-secant of the first and third quadrant can be mostly ascribed to those volatiles presenting high scores on both PC1 and PC2 (benzaldehyde (28), ethyl decanoate (32), geranyl acetone (39) and butyl butanoate (16)) or low scores on both components (propan-2-ol (5), ethanol (6), 3-methylbutan-1-ol (15) and 2-methylpropanoic acid (30)). A more generalized conclusion can be made looking at chemical families highlighted in Figure 4.3(B). Carbonyl compounds, others and esters, are both characterized by positive loading values on PC2 while most of the compounds belonging to alcohols lay on the third quadrant of PC1-PC2 plane (where PF class is localized). Chemical family of acids seems to be more dispersed. Moreover, looking at the loading behavior in Figure 4.3(C) and Figure 4.3(D), it can be added that, whereas

PC1 seems to describe the overall content of volatiles in samples, PC2 provides more information about content of alcoholic compounds which largely prevail in PF cheeses (see Table 4.2).

4.2.4 Discrimination of Pecorino cheese samples

The three typical Pecorino cheeses were differentiated using two discriminant approaches and the results were compared. LDA and PLS-DA models were calibrated on the same training set, consisting of 10 PF, 10 PS and 11 PR samples, while validation of the classification models was finally performed on an external set collecting the remaining samples (4 PF, 6 PS and 6 PR). To ensure representativeness of both data sets, sample splitting between training and validation sets was performed using Duplex[235] algorithm.

4.2.4.1 LDA model

Stepwise LDA was performed on the data matrix subjected to logarithmic transformation in order to obtain an optimal ratio between number of variables and objects. The most discriminant variables were iteratively selected according to the Wilks' lambda trend by the sequential addition of predictors; each addition/elimination of a further variable is evaluated in a leave-one-out cross-validation procedure. The performances in calibration, cross-validation (venetian blinds, five cross-validation groups) and prediction of the final 10-variable LDA model are reported in Table 4.3 in terms of correct classification rate (CCR%).

Table 4.3. Classification performances in terms of correct classification rate (CCR%) in calibration, cross-validation and prediction provided by the LDA and PLS-DA models.

	CCRs%, LDA			CCRs%, PLS-DA		
	PF	**PS**	**PR**	**PF**	**PS**	**PR**
computed classes	100	100	100	100	100	100
cross-validation (training set)	90.0	100	100	100	100	100
prediction (test set)	75.0	83.3	100	75.0	100	83.3

[235] Snee, R.D. *Technometrics*. **1977**, *19*, 415-428.

The 10 selected variables are reported in Figure 4.4.

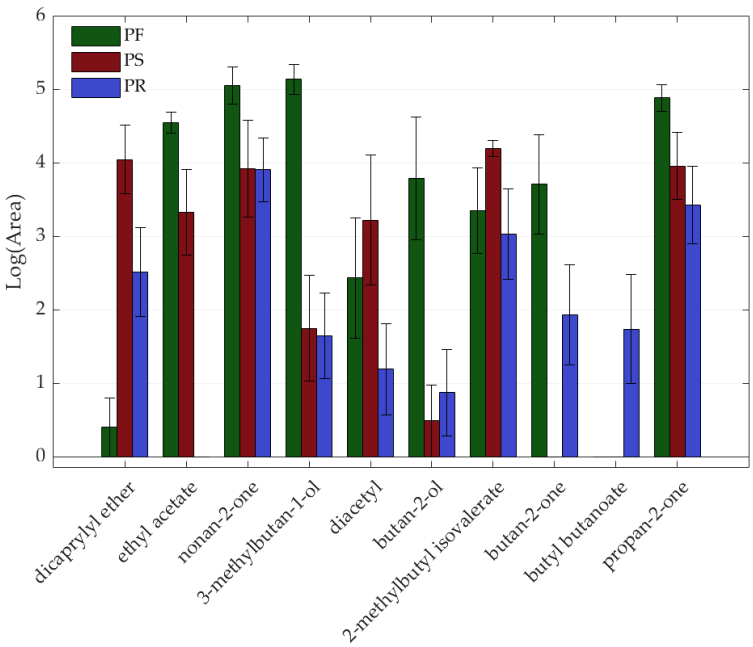


Figure 4.4. Mean values of the logarithm-transformed peak areas (with related standard errors) of the volatiles selected by stepwise-LDA for each Pecorino class.

The cross-validated calibration model erroneously assigned only one PF sample, while two classification errors, concerning one PS and one PF sample, occurred in prediction. The resulting values of non-error rates in prediction indicate a good generalization ability of the LDA model, taking into account that prediction errors in percentage terms are amplified because of the relatively low number of external objects. Calibration and validation samples are graphically displayed in Figure 4.5, where the objects (Figure 4.5(A)) and the weights of each LDA descriptor (Figure 4.5(B)) are projected in the plane of the first two canonical variables (CVs).

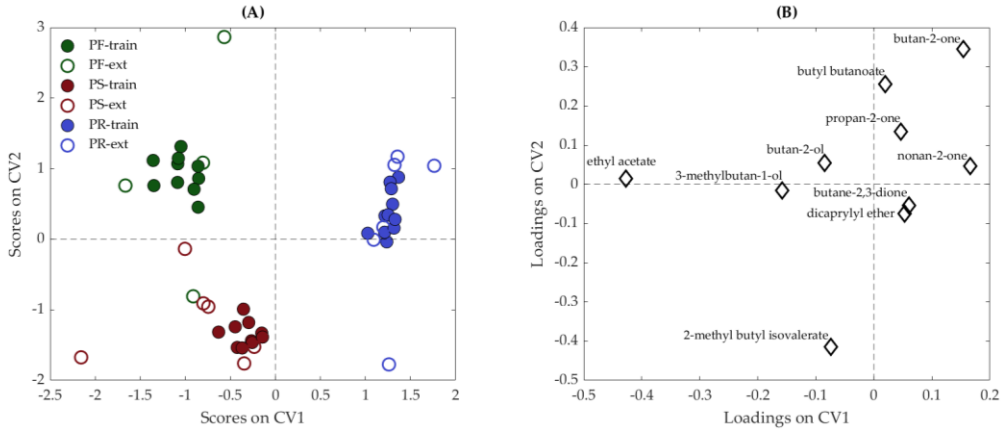


Figure 4.5. LDA Analysis. (A) Scores plot. External samples are represented by empty symbols. (B) Loadings plot.

In agreement with LDA performance reported in Table 4.3, Figure 4.5 reveals a good separation among the three classes in the CV1-CV2 space. Moreover, the external samples generally fall within the subspace of the calibration samples belonging to the same class, except for the misclassified samples. The PF and PR categories can be discriminated from the PS class along CV2, while CV1 describes the separation between PR samples and all the others. 2-methyl butyl isovalerate, butan-2-one and butyl butanoate seem to be the most influent variables in the discrimination of PF+PR and PS class while ethyl acetate, butan-2-one, nonan-2-one and propan-2-one are the most important variables in the separation of PF+PS and PR classes.

4.2.4.2 PLS-DA model

PLS-DA model parameters (the optimal number of latent variables and the most suitable data pre-treatment) were established on the basis of a venetian blinds cross-validation with five cancellation groups using the error rate as the indicator of the prediction performance. The pre-processing approach applied on the data matrix (previously subjected to logarithmic transformation) and ***Y*** response was autoscaling. The final model was calculated with eight latent variables explaining 70.0 and 65.5% of variance in ***X*** and ***Y*** blocks, respectively; classification rates are shown in Table 4.3. The PLS-DA model provided in external validation a root mean square error (RMSEP) of 0.40 and a 100% of non-error rate for class PS, and it misclassified only one object for each of the other two categories. The PLS-DA results are graphically shown in Figure 4.6, where the samples are projected onto the first two latent variables (LV1 and LV2) space.

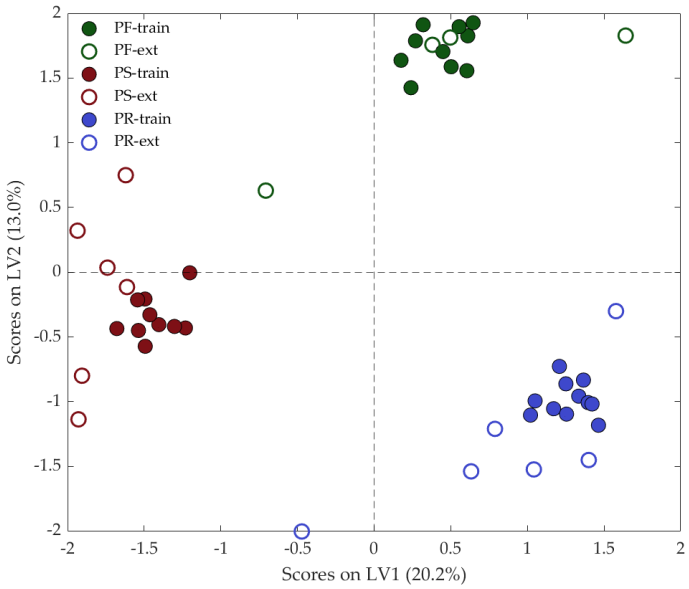


Figure 4.6. PLS-DA Analysis. Score plot. External samples are represented by empty symbols.

According to the results reported in Table 4.3, the scores plot reveals a clear separation among the calibration samples belonging to the three classes. In particular, LV1 discriminates the samples belonging to PS class from the other two categories and the PF category is discriminated from PS+PR group along LV2. All the Pecorino cheese test samples are relatively close to the respective calibration samples, except the two misclassified Pecorino cheeses (one PF, the same sample also erroneously classified in LDA, and one PR sample) that fall away from their class centroids and are closer to the PS class. A dept insight into the PLS-DA model can be given calculating the Variable Importance in Projection[152] (VIP) indices for each variable (reported in Table 4.4).

Table 4.4. VIP values, orderdered according to the decreasing importance in PLS-DA classifcation. The variables with VIP>1 are marked in bold.

ID	Variable	VIP	ID	Variable	VIP
4	**butan-2-one**	**1.65**	18	3-hydroxybutan-2-one	0.91
11	**pentan-2-ol**	**1.62**	39	geranylacetone	0.87
3	**ethyl acetate**	**1.61**	27	2-propylpentan-1-ol	0.86
35	**dicaprylyl ether**	**1.48**	26	acetic acid	0.85
29	**propanoic acid**	**1.45**	24	ethyl heptanoate	0.84
15	**3-methylbutan-1-ol**	**1.22**	31	butanoic acid	0.80
5	**propan-2-ol**	**1.16**	10	ethyl butanoate	0.79
32	**ethyl decanoate**	**1.15**	8	butane-2,3-dione	0.77
20	**heptan-2-ol**	**1.14**	33	3-methylbutanoic acid	0.76
9	**butan-2-ol**	**1.10**	23	nonanal	0.76
16	**butyl butanoate**	**1.09**	2	3-octene	0.75
7	**pentan-2-one**	**1.07**	28	benzaldehyde	0.75
6	**ethanol**	**1.02**	38	hexanoic acid	0.69
30	**2-methylpropanoic acid**	**1.02**	21	anisole	0.69
36	2-phenylpropan-2-ol	0.97	14	limonene	0.62
19	2-methyl butyl isovalerate	0.96	34	pentanoic acid	0.61
12	butan-1-ol	0.93	17	ethyl hexanoate	0.55
1	propan-2-one	0.93	37	unknown#1	0.51
25	ethyl octanoate	0.92	13	heptan-2-one	0.46

As customarily done, variables presenting a VIP index higher than 1 were considered contributing the most to the model. Accordingly, 14 compounds (butan-2-one, pentan-2-ol, ethyl acetate, dicaprylyl ether, propanoic acid, 3-methylbutan-1-ol, propan-2-ol, ethyl decanoate, heptan-2-ol, butan-2-ol, butyl butanoate, pentan-2-one, ethanol and 2-methylpropanoic acid) were considered the most relevant in the discrimination of the Pecorino samples. Of these, six were also selected by stepwise LDA (butan-2-one, ethyl acetate, dicaprylyl ether, 3-methylbutan-1-ol, butan-2-ol and butyl butanoate).

A point deserving attention is the relatively large variability of PF samples, evident not only in LV1-LV2 but also in PC1-PC2 and CV1-CV2 projections, which apparently contrast with the restricted production site of PF samples compared to the large production areas of both PS and PR cheeses. The reason of the high inner-variability in class PF (greater than PS and PR) can be found in the productive process rather than on the dimension of the production area. In fact, this kind of Pecorino is generally produced in a reduced number of small dairies, where the cheese-making process is in a small-scale, and it is carried out manually, with a minimum of automated tools. This, on one side, provides the uniqueness that distinguishes this Slow Food Presidia product, but, on the other, leads to an increased variability in the characteristics of the loafs which is reflected on the wide inner-class variance observed here. Contrarily, PS and PR cheeses, in order to maintain the PDO label, are forced to strictly respect specific characteristics. These cheeses are often produced in medium-sized factories, which exploit machinery in the production phase achieving high reproducibility of the productive procedure and

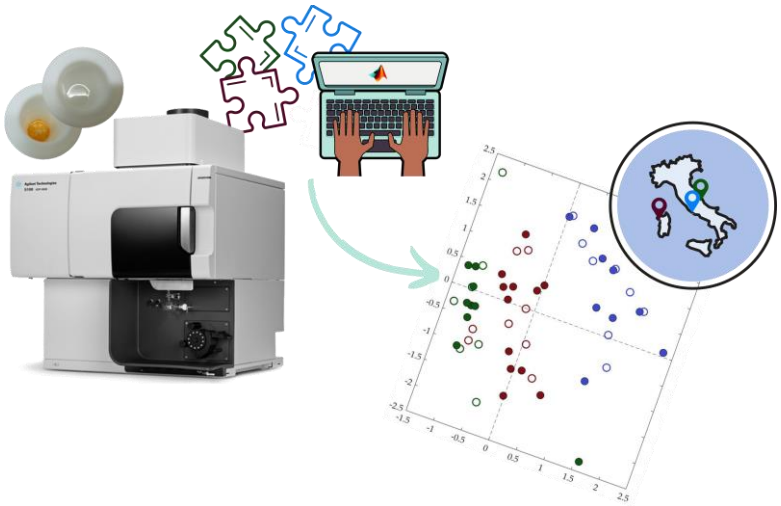
high co-conformity among loafs. Consequently, the inner-class variability for samples belonging to these two categories are reasonably narrow.

4.3 Conclusions

The goal of the present work was to build an analytical protocol for the characterization and discrimination of three typical and certified Pecorino cheeses produced in Italy with different and peculiar technologies: Pecorino di Farindola (Slow Food Presidium), Pecorino Sardo (PDO) and Pecorino Romano (PDO). For this purpose, 47 Pecorino samples were analyzed by means of HS-SPME/GC–MS and finally classified by two discriminant approaches, LDA and PLS-DA. The samples collected from several productors belonged to, at least, two manufacturing cycles and were characterized by different ripening times in order to verify the robustness of the classification models. The difficulty in finding samples surely belonging to the studied labels (PDO for PS and PR, Slow Food Presidium for PF) was the cause of the limited number of objects in both the calibration and validation sets. This condition led to a greater impact of each misclassification on the overall performance of the models. Nevertheless, regardless the natural variability in Pecorino aroma, related to ripening time and seasonal effects, the analysis of volatile compounds coupled with discriminant classifiers provided satisfactory results, correctly assigning all the calibration samples and most of the external objects. In fact, both approaches provide a total classification rate (in external validation) of 87.5%, corresponding to only two test objects erroneously assigned. Of these, one (belonging to class PF) is misclassified by both classifiers. Consequently, both the proposed strategies (aroma profiling coupled with stepwise LDA or with PLS-DA) were effective in the indication of the Pecorino origin and cheesemaking procedure.

Chapter 5

Multi-Elemental Composition Data Handled by Chemometrics for the Discrimination of High-Value Italian Pecorino Cheeses[67]



5.1 Materials and Methods

5.1.1 Pecorino cheese samples and chemicals

Fifty-three (53) samples were available for the discrimination study (16 PF, 20 PS and 17 PR). More detailed information on the investigated Pecorino samples is reported in Section 4.1.1 of Chapter 4.

A multi-element TraceCERT® standard solution for ICP (Fluka Analytical, Sigma Aldrich) containing Ba (at 40 mg/L), Fe and Zn (both at 100 mg/L) and mono-element TraceCERT® certified standards for AAS of Ca, K, Mg, Na and P (1000 mg/L in nitric acid), purchased from Sigma Aldrich, were used to prepare an analytical calibration curve with six standard solutions (by dilution in 50 and 100 mL polymethylpentene volumetric flasks) plus a blank consisting of ultrapure water (18.2 MΩ cm resistivity at 25°C) obtained with a water purification system MilliQ (Millipore, Germany). Hydrogen peroxide solution (≥ 30% *w*/*w*) for ultra-trace analysis from Sigma-Aldrich (St. Louis, MO, USA) and suprapure nitric acid (65% *w*/*w*) from Merck KGaA (Darmstadt, Germany) were used to digest the samples.

5.1.2 Sample preparation and Microwave-Assisted Digestion

Pecorino cheese samples were dried in a Büchi TO-50 (Büchi Labortechnik AG, Flawil, Switzerland) drying glass oven at 75°C for 24 h under moderate vacuum and successively ground in a mortar (cleaned after each sample preparation with an aqueous solution at 2%*v*/*v* of HNO_3 and ultrapure water) before the digestion procedure. Aliquots of 0.3 g of cheese were transferred into Teflon vessels and, after the addition of 5 mL of hydrogen peroxide and 5 mL of nitric acid, were digested in an Ethos One (Milestone, Bergamo, Italy) microwave oven at constant power (1000 W). Complete mineralization was obtained in 40 min at the constant temperature of 180°C (reached in 10 min). After cooling at room temperature, the obtained solutions were diluted with ultrapure water into 25 mL PMP volumetric flasks. Fortified samples were prepared in the same way to check the trueness of the analytical method, in the absence of a Certified Reference Material for this matrix. Aliquots to be used in the preparation of both genuine and enriched samples were collected by the same slice of cheese, accurately homogenized. Recovery studies were performed separately for micro- (Ba, Fe and Zn) and macro-elements (Ca, Na, P, K and Mg) taking into account of the sample mass (0.3 g) and the different dilutions used in the measurements. The spiking levels were chosen according to native concentrations of the detected elements in Pecorino cheese (determined in preliminary analyses) i.e., 2.67 $\mu g/g_{dry}$ of Ba, 6.67 $\mu g/g_{dry}$ of Fe and Zn (dilution to 25 mL) and 83.3 $\mu g/g_{dry}$ of Mg, 166.7 $\mu g/g_{dry}$ of K, 833.3 $\mu g/g_{dry}$ of Na and P, and 1666.7 $\mu g/g_{dry}$ of Ca (dilution to 100 mL).

5.1.3 ICP-OES Analysis

The digested samples were transferred into polypropylene vials for the simultaneous analysis of eight elements (Ba, Ca, Fe, K, Mg, Na, P and Zn) by means of an Iris Intrepid ER/S Thermo-Elemental (ThermoScientific, Waltham, MA, USA) ICP-OES spectrometer equipped with an Echelle grating optical system and a charge injection device (CID) solid-state detector. A Timberline II (ThermoScientific) autosampler ensured the introduction of the samples with a flow rate of 1.85 mL/min, controlled by a peristaltic pump. Each sample was conveyed in a concentric pneumatic nebulizer, where it was nebulized (nebulizer gas flow of 0.6 mL/min) into the argon plasma, connected to a cyclonic spray chamber and eventually analyzed in radial torch-reading configuration using the operating conditions recommended by the manufacturer, i.e., radiofrequency power of 1.202 kW, coolant gas flow of 12 L min^{-1} and auxiliary gas flow of 0.5 L min^{-1}. Ultra-High Purity 5.0 Grade Argon (99.999% pure argon) was used in the spectrometric analysis. The emission lines that provided the maximum signal to noise ratio and minimum spectral interferences were selected and a manual background correction of the emission intensity was performed. Each measurement, both on standards and digested samples, was performed in four replicates and the mean value was taken. Matrix effects, possible non-spectral interferences and instrumentation drift were

monitored using a 200 μg/L Yttrium solution as Internal Standard (line 324.228 nm). Cleanliness of the introduction system and absence of memory effects were controlled by the analysis of one standard solution (0.024 μg/mL for Ba, 0.06 μg/mL for Fe and Zn, 10 μg/mL for Mg, 20 μg/mL for K, and 30 μg/mL for Ca, Na and P) followed by a blank every six samples. The Pecorino samples were analyzed in a random order.

5.1.4 Statistical Analysis

The data matrix consisting of the concentration values of eight elements (Ba, Ca, Fe, K, Mg, Na, P and Zn) for each sample was preliminarily handled by one-way ANalysis Of VAriance (ANOVA). Means concentrations of the above elements in the Pecorino samples grouped according to geographical origin were compared to check for significant effects within multi-elemental composition. In addition, a multiple pairwise comparison of the observed means with LSD criterion was performed to highlight couples for which the content of each element differs. The problem of classifying the samples according to their geographical origin and cheesemaking protocol was coped with PLS-DA. The Matlab (R2020b; The Mathworks, Natick, MA) function for PLS-DA here employed can be freely downloaded from the RomeChemometrics website[221]. For further details about the theory of the mentioned Chemometric techniques, the reader can refer to Chapter 3.

5.2 Results and Discussion

5.2.1 Validation of ICP-OES Analysis of Pecorino cheeses

The ICP-OES method developed for the multi-elemental analysis of Pecorino cheeses was validated according to the EURACHEM[236] guidelines. The figures of merit associated to this method are summarized in Table 5.1. Linearity and sensitivity were derived for each element from the related analytical calibration curve as determination coefficient (R^2), and Limit of Detection (LOD) and quantification (LOQ), respectively. The observed R^2 values are greater than 0.9995, except for Fe (R^2 = 0.9921), ensuring the good linearity for all the elements in the working range investigated. LOD and LOQ values were both determined according to the BEC[237] (Background Equivalent Concentration) concept. The relations $LOD=3 \cdot RSD_{blank} \cdot BEC/100$ and $LOQ = 10 \cdot RSD_{blank} \cdot BEC/100$ were used, where RSD_{blank} is the Relative Standard Deviation (n=10) of the blank and BEC is equivalent to the concentration producing twice the intensity of the background. Both LOD (from 0.05

[236] Eurachem. Eurachem Guide: The Fitness for Purpose of Analytical Methods–A Laboratory Guide to Method Validation and Related Topics. Retrieved from www.eurachem.org (last access January 2022).
[237] Thomsen, V.; Roberts, G.; Burgess, K. *Spectroscopy*. **2000**, *15*, 33-36.

to 5.64 μg/g_{dry}) and LOQ (from 0.18 to 18.80 μg/g_{dry}) are largely below the mean native levels of each target element in Pecorino, even for Fe, which presents a content in the sample five times the corresponding LOQ values. Therefore, sensitivity is good enough to ensure a reliable determination of the detected elements in the real samples. Accuracy was evaluated by a spike recovery analysis because a Certified Reference Material for cheese matrix was not available. Six genuine samples and six fortified ones were analyzed (see details in Section 5.1.3) to estimate the mean recoveries (R (%)) of the target elements (Table 5.1). Precision of the ICP-OES method was estimated by the Relative Standard Deviation (RSD (%)) determined in six procedural replicates. Comparison of the above two parameters with the acceptable values reported in the specialized literature [238] confirms that the eight detected elements can be determined with satisfactory accuracy and precision. As reported in Table 5.1, the elements Ba, Fe and Zn (1–50 μg/g range of concentrations) showed R(%) values from 83 to 102% (recovery benchmark: 80-110%) and RSD% values from 1 to 12% (AOAC threshold 7.3-11%, Horwitz threshold 11.3-16%); K and Mg (0.6 μg/g-2 mg/g range of concentrations) were detected with mean recovery percentages of 98 and 99% (recovery benchmark: 90−107%) and RSD% values from 1 to 2% (AOAC threshold 5.3%, Horwitz threshold 8%) and the elements P, Ca, and Na (8-20 mg/g range of concentrations) with mean recovery of 99-100% (recovery benchmark: 97-103%) and RSD% values from 1 to 2% (AOAC threshold 2.7%, Horwitz threshold 4%).

Table 5.1. Working linear range (WLR), determination coefficient (R^2) and limit of detection (LOD) and quantification (LOQ) related to the analytical calibration curve of the elements (with corresponding detection wavelength, λ) determined by ICP-OES. Mean recovery (R(%)) and relative standard deviation (RSD(%)) observed in the analysis of the enriched samples (n=6).

Element	λ (nm)	R^2	Working linear range (μg/mL)	LOD (μg/g_{dry})	LOQ (μg/g_{dry})	R(%)	RSD(%)
Ba	233.527	0.9998	0.008-0.08	0.05	0.18	97	7
Ca	215.887	0.9997	10-100	3.38	11.27	99	2
Fe	259.940	0.9921	0.02-0.20	0.22	0.73	83	12
K	766.491	0.9996	5-50	1.43	4.78	98	1
Mg	280.271	0.9997	1-30	1.01	3.35	99	2
Na	588.995	0.9996	10-100	5.64	18.80	99	1
P	213.618	0.9998	10-100	2.80	9.33	100	2
Zn	202.548	0.9975	0.02-0.20	0.24	0.81	102	1

[238] a) Huber, L. *Validation and Qualification in Analytical Laboratories* (2nd edition), London, UK: Informa Healthcare, **2007**; b) Gustavo González, A.; Ángeles Herrador, M. *TrAC-Trends Anal. Chem.* **2007**, *26*, 227-238.

5.2.2 Exploratory Analysis of ICP-OES data

Table 5.2 displays the mean concentrations (referred to the sample dry weight) of the eight detected elements in the three Pecorino cheese varieties along with the related standard deviations. Analysis of Variance (ANOVA), carried out to assess which means differ significantly, reveals that the mean content of Ca, Mg and P is not significantly affected by the cheese origin, whereas, on the opposite side, the mean concentrations of Ba, K and Na exhibit highly significant differences within the three groups of Pecorino cheeses. Successive pairwise comparison by least significant difference (LSD) test reveals that Ba is the only element that exhibits a significantly different mean concentration among the three Pecorino varieties, while the mean contents of other elements (Fe, K, Na, and Zn) are significantly different only between specific Pecorino pairs (Table 5.2).

Table 5.2. Mean concentrations and related standard deviations of the detected elements in the Pecorino cheese samples. Significance (*p*-value) of difference in the means determined by one-way ANOVA and list of the significantly different class pairs according to LSD test.

	mean±SD				
Element	**PF**	**PR**	**PS**	**ANOVA (p value)**	**LSD (0.05 significance level)**
Ba*	1.2±0.7	3.5±1.1	2.7±0.8	$<10^{-4}$	PF-PS; PS-PR; PF-PR
Ca$	12±3	13±3	13±2	0.307	-
Fe*	2.4±1.7	3.8±1.8	2.5±1.3	0.028	PF-PR; PS-PR
K$	1.7±0.3	1.1±0.3	1.5±0.4	$<10^{-4}$	PF-PR; PS-PR
Mg$	0.65±0.19	0.73±0.16	0.70±0.12	0.333	-
Na$	11±5	21±5	10±2	$<10^{-4}$	PF-PR; PS-PR
P$	8±2	9.0±1.7	8±2	0.691	-
Zn*	40±12	49±15	48±11	0.071	PF-PS; PF-PR

* (mean±SD) $\mu g/g_{dry}$
$ (mean±SD) mg/g_{dry}

The results of PCA exploration, performed on the auto-scaled data matrix, are displayed in Figure 5.1 showing the projection of both samples and variables in the plane of the first two PCs.

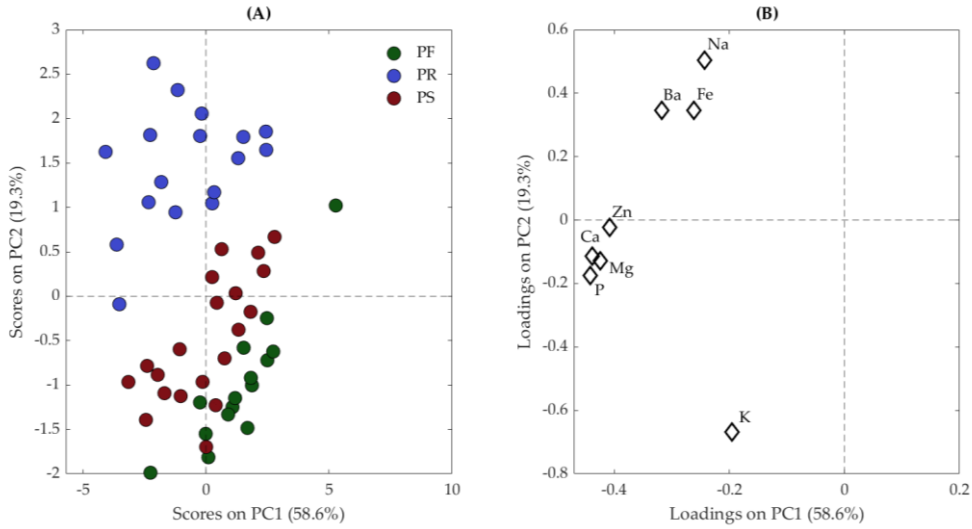


Figure 5.1. PCA Analysis. Projection of Pecorino samples (A) and variable loadings (B) on the first two principal components (PC1 and PC2).

It can be observed that the samples belonging to each of the three Pecorino varieties are distributed approximately along the bisector of the graph. Therefore, this direction describes the variability internal to each kind of Pecorino that is expected to take origin from the diversity in the ripening time, cheesemaking process and location of the production site within the deputed territory. In this regard, the variability of PF samples along this direction is comparable to that of the other two varieties, despite the first cheese is produced in a relatively small territory rather than on a regional (PS) or extra-regional (PR) scale. PF is also produced in a reduced number of small dairies by small-scale cheesemaking processes generally carried out manually. This, from one side, confers unicity to the product, as recognized by Slow Food Presidia, but, on the other, may increase the inner-class variability. PS and PR cheeses, by contrast, are often produced in medium-sized factories with the support of mechanical tools, that in compliance with the PDO specifications, ensures a better control of the cheesemaking process and, therefore, should reduce the dissimilarity among different cheese lots.

The Pecorino samples grouped according to the origin along PC2, although some of the three subgroups partially overlap. The PR samples, all having positive PC2 scores, are separated from PS and PF individuals, which fall at negative values of this component and whose discrimination in the PC1-PC2 plane is incomplete. The investigation of the loadings makes apparent the sequential location of PF, PS, and PR clusters along PC2 is mainly associated to the varying content of Na, Ba, Fe, and K. Looking at the PCA score and loading plots displayed in Figure 5.1, it can be highlighted that PC1 seems to describe the overall mineral content in the cheeses, whereas the sequential location of PF, PS, and

PR samples along PC2 is mainly associated to the varying content of K (negative loadings on this component) and Na, Ba, and Fe group (all having positive loadings). Such behavior cannot be interpreted in a straightforward way because information about the processing-variables, potentially affecting the mineral composition of the final product, is not always explicitly reported in the production guidelines. Nevertheless, the results here obtained seems to be consistent with those of a previous investigation dealing with the role of raw milk source, type of rennet, salting-procedure and ripening time on the mineral composition[239]. In general, cheeses produced with sheep's milk[240], higher ripening time and the use of animal rennet[241] present a higher overall content of minerals. Moreover, the type of rennet, that influences moisture, pH, fat and protein content[182b], is expected to affect also the mineral composition. Concerning salting procedure, PR and PS are produced using both immersion (wet) and sprinkling (dry) salting, whereas only dry-salting is applied in PF production. It is reported that immersion in a brine bath is the technique resulting in the lowest level of potassium[239], in agreement with the negative scores of most of the PF samples on PC2. On the contrary, the level of Fe is reported to be higher in wet-salted cheeses, which agrees with the observed anti-correlation of K and Fe variables. Na contents are higher in longer ripening time cheeses, which justifies the location of PR, commonly known as a *salty* cheese, at positive PC2 scores. Multi-elemental composition is also affected by many other factors, i.e., animal breeding[242], physiological and environmental conditions of the animals[243], and pH value during cheese processing (influencing mineral losses), but these effects cannot be invoked here because the above conditions are not described in the production guidelines of the investigated cheeses.

As previously discussed, the variability in the multi-elemental composition within the three Pecorino classes can be also affected by the fact that the cheeses are produced in two different manufacturing cycles and, in the case of the PS and PF groups, at two different levels of ripening. Although not reported in Figure 5.1 for clarity, the PS samples belonging to the *dolce* category are mainly characterized by higher scores on PC2, whereas discrimination of the PF samples according to the ripening time was less evident. Concerning the effect of the manufacturing cycle, no differentiation of the PS samples was observed, while a clear and a less perceptible differentiation of PR and PF samples, respectively, was detected in the PC1-PC3 plane.

---

[239] Moreno-Rojas, R.; Cámara-Martos, F.; Sánchez-Segarra, P.J.; Amaro-López, M.Á. *Int. J. Dairy Technol*. **2012**, *65*, 594-602.
[240] Raynal-Ljutovac, K.; Lagriffoul, G.; Paccard, P.; Guillet, I.; Chilliard, Y. *Small Rumin. Res*. **2008**, *79*, 57-72.
[241] Patel, H.G.; Upadhyay, K.G.; Pandya, A.J. *Asian J. Dairy Res*. **1991**, *10*, 217-222.
[242] Adams, R.S. *J. Dairy Sci*. **1975**, *58*, 1538-1548.
[243] Anon, T. *Int. Dairy Feed. Bull*. **1981**, *140*, 5-19.

5.2.3 PLS-DA geographical classification of Pecorino

Prior to the creation of the classification model, the Duplex[235] algorithm was applied on each class separately to divide the samples into a training and a test set. Of the 53 Pecorino cheese samples, 30 (9 PF, 12 PS and 9 PR) were selected as the training set, while the remaining 23 (7 PF, 8 PS and 8 PR) were left out for the external validation of the model (test set). PLS-DA was optimized applying Linear Discriminant Analysis (LDA) on the calculated/predicted-$\boldsymbol{Y}$ responses in a 5-fold cross-validation procedure and exploring the evolution of classification error by increasing the number of latent variables. The classifier was tested on raw- and logarithmic scaled matrix, but no classification improvement was observed after pre-processing of the data; regardless the pre-treatment used, data was auto-scaled prior to calculations. The PLS-DA model with three latent variables, which explained 85% of variance on $\boldsymbol{X}$-block and 95% on $\boldsymbol{Y}$-block, was eventually retained. This model performed very well on the training set, showing 93.3% accuracy in cross-validation, corresponding to the misclassification of 1 PF and 1 PR sample. A similar accuracy was observed in prediction (91.3%) proving a great stability and balance between the training and test set. All the external samples belonging to PF and PR classes were correctly assigned, while only two PS samples were misclassified. A graphical representation of the results of the PLS-DA analysis is provided in Figure 5.2.

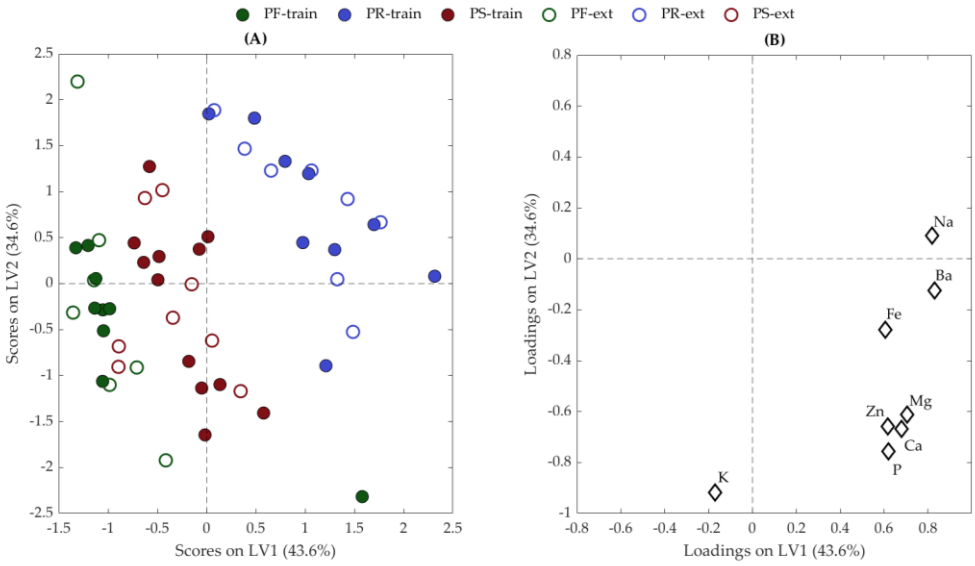


Figure 5.2. PLS-DA Analysis. Projection of Pecorino samples (**A**) and variable loadings (**B**) on the first two latent variables (LV1 and LV2). Full and open circles identify the training and the test samples, respectively.

A further inspection of the Variable Importance Projection (VIP) scores allowed the identification of the variables contributing the most to the model, according to the *greater-than-one* criterion. VIP analysis identified only three significant predictors, i.e., the elements Ba, K and Na. Afterwards, a novel PLS-DA model was built on the reduced data set (i.e., exploiting only Ba, K and Na); nevertheless, the classification performance provided by this further model was the same as the complete model. The variable selection only reduced the intra-class variances in the space of the latent variables, in line with the considerations reported in Section 5.2.2.

The same three varieties of Pecorino cheese were previously classified by PLS-DA using the volatile profiles (described in the previous chapter)[66], which is worth comparing to the present study. The PLS-DA model described in this Section exhibits a good prediction ability, being capable of correctly classifying 91% of the external samples (21 out of 23), despite only few elements (K, Na and Ba) are significant. As displayed by ANOVA and LSD tests reported in Table 5.2, higher Na and lower K contents are typical of PR cheese, while a significant increase of Ba concentration can be observed in the three Pecorino varieties following the order PS<PF<PR. The predictive performance of the PLS-DA model based on the volatile profiles, although built on 14 components of the cheese aroma, was not as good as that provided by the above three significant elements. Moreover, interpretation of the PLS-DA model based on the volatiles is particularly difficult because of the many variables affecting the complex process of cheese aroma generation. Preliminary exploratory PCA analysis conducted on the volatile profiles revealed a separation of PF samples and a partial overlapping of PS and PR individuals. This suggests that geographical origin of milk has a relevant role in the differentiation of the three varieties according to the aroma, whereas the different mineral composition of the three varieties is also dependent on cheese-making technology, such as the salting mode.

5.3 Conclusions

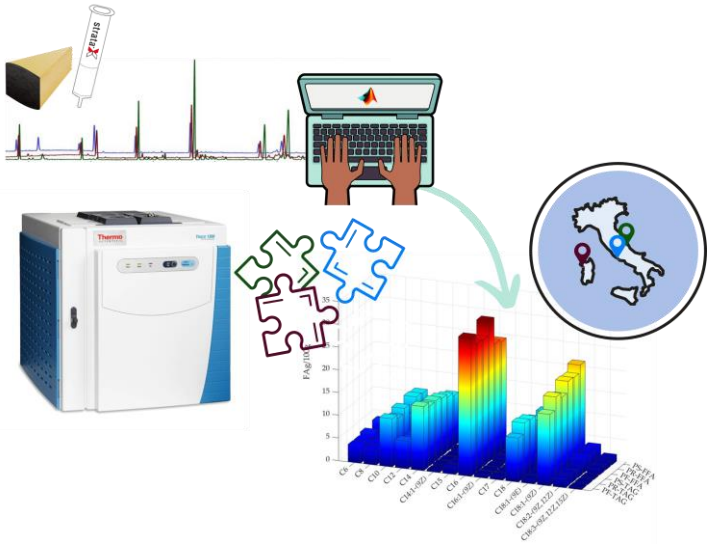
From the inspection of the outcomes of the PCA and PLS-DA models illustrated in the previous sections, it is quite evident the diverse classes of Pecorino present noticeable differences among one another. As expected, the divergencies initially highlighted by PCA were confirmed by the PLS-DA model. As described, these discrepancies are not based solely on the diverse origins of the cheeses, but also on the different procedures followed for their preparation. The elemental analysis allowed revealing macroscopic differences among the concentrations of the 8 investigated elements; nevertheless, the VIP analysis opened up to a more refined interpretation of which variables contribute the most to the classification model. In particular, in complete agreement with the outcome of the ANOVA, it became apparent the discrimination is mainly due to Ba, Na, and K. The inspection of the PCA-loadings plot revealed that, of these, the first two are

found at higher concentrations in PR samples than in the other two classes; on the contrary, K is particularly high in PS and PF, whereas is anticorrelated with PR.

As far as the predictive aspect of the classification model is concerned, it is evident that the PLS-DA model is robust and reliable, and it erroneously classifies only two test samples, belonging to class PS. A more in-depth investigation of these individuals has shown that they are both Pecorino *dolce*, i.e., soft-ripening; this aspect surely influenced their mineral composition and, consequently, their class-assignment.

Chapter 6

Fatty acid composition of the triglyceride and free fatty acid fractions in typical Italian Pecorino cheeses



6.1 Materials and Methods

6.1.1 Pecorino cheese samples and chemicals

Fifty (50) samples, 18 PS, 17 PR and 15 PF, were available for this investigation.

More detailed information on the investigated Pecorino samples is reported in Section 4.1.1 of Chapter 4.

The solid phase extraction (SPE) cartridge Strata® NH2 500mg/6mL (Phenomenex, Torrance, CA, USA) based on moderately polar, silica-based bonded phase (55 µm, 70Å) with weakly basic surface (aminopropyl) sorbent was employed to achieve an effective concentration of the analytes. FAME mixtures Grain FAME Mix (Supelco, Bellafonte, PA, USA) and C4-C24 even carbon satured FAMEs (Supelco, Bellafonte, PA, USA) were used in the GC analysis to identify the peaks by comparing the retention times.

6.1.2 Fat extraction

Fatty acid (FA) content in Pecorino cheeses was determined after modified fat extraction according to the method described by de Jong and Badings[244]. Briefly, a mixture of 1 g of grated cheese, 2 g of anhydrous sodium sulfate (Na_2SO_4), and 0.3 mL of sulfuric acid (H_2S0_4, 2.5 M) solution was extracted three times with 3 mL ether/heptane (1:1, v/v) from Sigma-Aldrich (St. Louis, MO, USA). Each time the solution was clarified by short centrifugation and the upper solvent layer was added with 1 g of anhydrous Na_2SO_4 to adsorb residual water. This procedure achieved the extraction of free fatty acids (FFAs) and neutral lipids (mostly triacylglycerols, TAGs), then the isolation of the two fractions was performed using a weak anion-exchanger[245]. The SPE sorbent was conditioned by washing with 10 mL of heptane and equilibrated with 2 mL ether/heptane (1:1, v/v). The solvents were forced through the cartridge by means of a positive pressure at a rate of approximately 1 mL/min. The sample was passed through the cartridge by applying a moderate vacuum. The retained analytes were sequentially eluted: neutral lipids (triacylglycerols) by 10 mL of chloroform–2-propanol mixture (2:1, v/v) and FFAs by 2 mL of acidified (2% formic acid) ether. The elutes were collected and differently processed.

6.1.2.1 Triacylglycerol (TAGs)

The fatty acid methyl esters (FAMEs) were prepared by the rapid trans-methylation method under alkali-catalyzed conditions with potassium hydroxide according to the International Organization for Standardization[246]. Briefly, the elute obtained was added with 200 µL of methanolic potassium hydroxide solution (2 M) and the glass vial was shaken vigorously for about 30 seconds. After an initial cloudiness due to the separation of glycerol, the solution became clear. About 1 g of sodium hydrogen sulfate monohydrate ($NaHSO_4 \cdot H_2O$) was added to neutralize potassium hydroxide and prevent saponification of the methyl esters. The solution was shaken briefly and, after the salt had settled, the upper layer was transferred into a vial. The obtained solution was suitable for GC analysis.

---

[244] de Jong, C.; Badings, H.T. *J. High Resolut. Chromatogr*. **1990**, *13*, 94–98.
[245] Kaluzny, M.A.; Duncan, L.A.; Merritt, M.V.; Epps, D.E. *J. Lipid Res*. **1985**, *26*, 135–140.
[246] International Standardization Organization (ISO, Geneva, Switzerland). **2017**. BS EN ISO 12966-2:2017: *Animal and vegetable fats and oils-Gas chromatography of fatty acid methyl esters. Part 2: Preparation of methyl esters of fatty acids*. Retrieved from https://www.iso.org/standard/72142.html (last access January 2022).

6.1.2.2 Free Fatty Acids (FFAs)

FAMEs were prepared by boron trifluoride (BF_3) methylation procedure, avoiding the first alkali-catalyzed step or saponification (strictly necessary only in the case of the presence of TAGs), according to the AOAC Official Methods of Analysis[247]. Briefly, the sample (elute obtained in Section 6.1.2) and 5 mL of BF_3-methanol solution at 10% for GC derivatization (Supelco, Bellafonte, PA, USA) were introduced into a flask, the air was removed by flushing the flask with dry argon and the condenser was fitted to it. The reagents had been left to boil under reflux for 2 minutes and 1 mL hexane (≥95%) from Sigma-Aldrich (St. Louis, MO, USA) was added to the boiling mixture through the top of the condenser. After 2 minutes the flask was removed from the heat source and the reflux condenser was disconnected. A saturated sodium chloride solution was added until the liquid level of the mixture was brought up to the neck of the flask. After the separation of the two phases, the upper organic layer was added with anhydrous Na_2SO_4 to adsorb residual water. The so-obtained hexane solution was suitable for GC analysis.

6.1.3 GC-FID Analysis

FAME profile was determined using a Thermo Trace GC Ultra 7820 gas chromatograph system (Thermo Fisher Scientific, Waltham, MA, US) equipped with a flame ionization detector (FID). The separation was achieved using a 30mx0.25mmx0.25µm fused silica capillary column SP™-2380 (Supelco, Bellafonte, PA, USA). The oven temperature was maintained at 80°C for 2 min, increased to 180°C at 10°C/min, held at 180°C for 10 min, increased again at 10°C/min to 250°C and held at 250°C for 2 min. The SSL injector was kept at 250°C during the whole analysis and injections were performed in split mode with 25 mL/min split flow. The detector temperature was maintained at 250°C with 350 mL/min of air and 40 mL/min of hydrogen. Hydrogen was constantly supplied by GENius PF500 (FullTech Instruments, Rome, Italy) generator and used as carrier gas at a constant pressure of 0.5 bar.

---

[247] AOAC International. AOAC Official Method 969.33. *Fatty Acids in Oils and Fats. Preparation of Methyl Esters. Boron Trifluoride Method*. First Action **1969**.

6.1.4 Statistical Analysis

The peak areas of the identified components were collected analyzing the GC chromatogram using Xcalibur™ software (Thermo Fisher Scientific, Waltham, MA, US). The non-identified FAMEs by contrast were not included in statistical analysis to guarantee a chemical interpretation of the classification models. The resulting data-table was exported to the Matlab environment (R2020b; The Mathworks, Natick, MA), and, as suggested by the International Organization for Standardization[248], the composition of FAMEs was calculated using the correction factors (CFs) derived from the analysis of the above-mentioned mixture standards, i.e., experimental CFs are *ad-hoc* derived, instead of using the theoretical ones reported in literature. The mass fraction of each FAME in grams per 100 g fat was eventually converted in grams of original FA per 100 g fat. Two blocks of data are eventually available for each sample: i) g/100g of FAs from the TAG fraction (analysis of sample prepared as in Section 6.1.2.1); ii) g/100g of FAs from the FFA fraction (analysis of sample prepared as in Section 6.1.2.2).

To achieve the goal of classifying the samples according to their geographical origin and cheesemaking protocol by exploiting both the information, different strategies of MB-data analysis were employed, and the results compared with the ones obtained from PLS-DA analysis performed on one block at the time and taken as benchmark. The Matlab functions employed in the present work can be freely downloaded from the RomeChemometrics website[221].

6.2 Results and Discussion

6.2.1 Fat composition

A qualitative comparison among typical obtained chromatograms is shown in Figure 6.1. Generally, TAG-chromatograms are more intense and richer of peaks, but the overall profile seems to be very similar between TAG and FFA-fraction of the cheeses.

---

[248] International Standardization Organization (ISO, Geneva, Switzerland). **2015**. BS EN ISO 12966-4:2015. *Animal and vegetable fats and oils-Gas chromatography of fatty acid methyl esters. Part 4: Determination by capillary gas chromatography*. Retrieved from https://www.iso.org/standard/63503.html (last access January 2022).

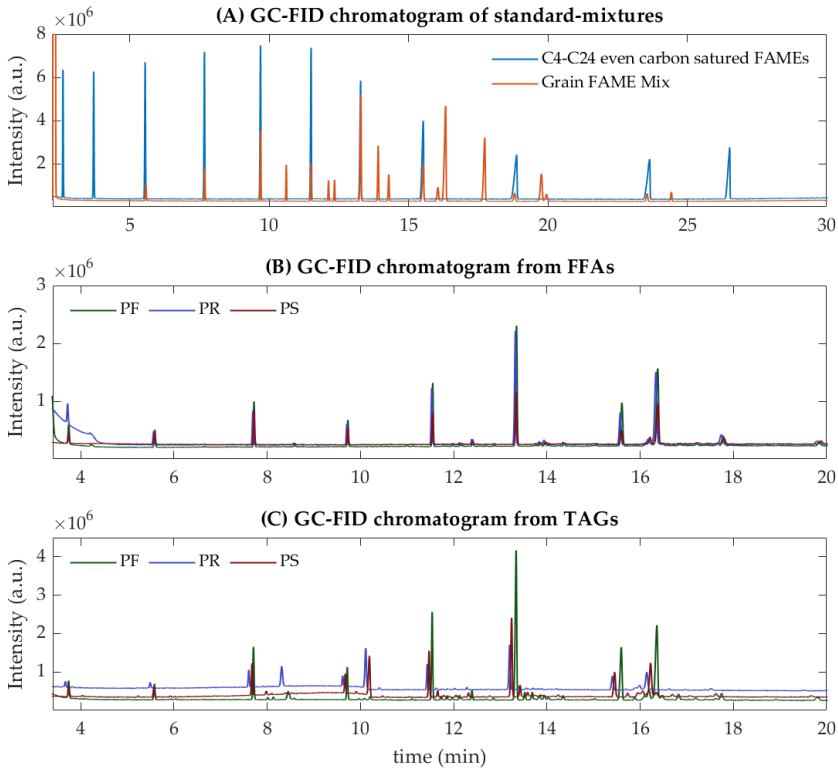


Figure 6.1. Comparison of GC chromatograms provided by PF, PR and PS samples of Pecorino cheeses collected from the (B) FFAs and the TAGs (C) fraction. The identification is performed looking at the retention times of the standards (A).

Tables 6.1 and 6.2 provided, respectively, the list of FAMEs with corresponding CFs present in the standard mixtures and the ones effectively found in the Pecorino cheese samples, together with their mean concentrations and related standard deviations. ANOVA was carried out to assess which means significantly differ and successive pairwise comparison by Tukey-Kramer revealed which groups among the three Pecorino varieties exhibit this difference.

The same information, i.e., the mean profiles of the FAs identified in TAG and FFA-fractions for the three Pecorino cheeses, can be graphically visualized looking at the Figure 6.2.

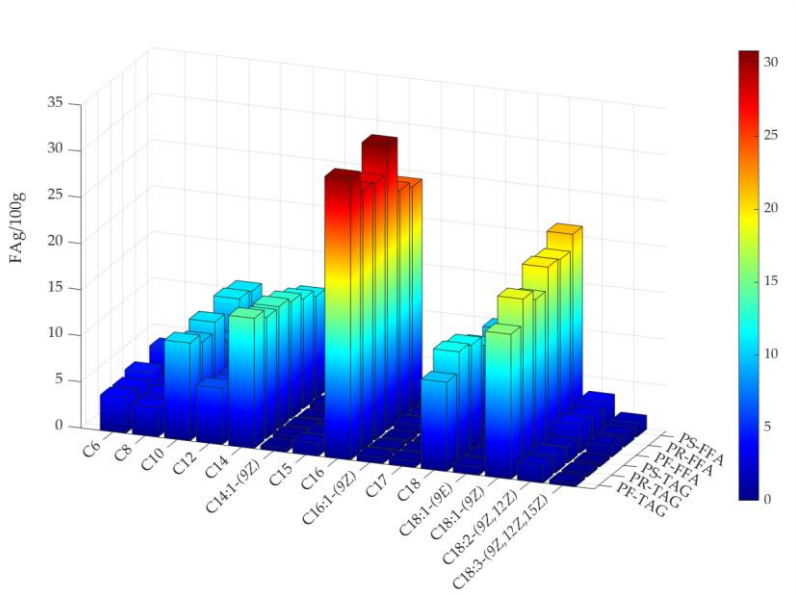


Figure 6.2. Average content in the triglyceride and free fatty acid fractions of PF, PR and PS. Each bar (i.e., FA) is colored according to its height (i.e., percentages in fat).

Due to the limited information regarding the individual FFA profiles of these cheeses, the results here obtained will be only partially related to previous knowledge in literature.

The most abundant FFAs found in Pecorino cheeses were palmitic (C16) and oleic acids (C18:1-*(9Z)*), followed by myristic (C14), capric (C10) and stearic (C18) acids like in the case of other cheeses[249]. The absence of butanoic acid and the lower percentages of short-chain FFA (C4-C10), apart from their volatility, can be attributed to the fact that part of these fatty acids released during the initial stages of ripening is converted into methyl ketones and other compounds[250]. Low levels of these short chain fatty acids can also indicate the low lipolytic action of the natural microflora of the cheese. The high percentages found for the capric acid (C10) are consistent with the literature; ewes'- and goats'-milk cheeses had indeed high percentage levels of C10 (around 10%), as compared to lower levels of around 4.5% in the cow milk cheeses[251].

---

[249] a) Mallatou, H.; Pappa, E.; Massooras, T. *Int. Dairy J.* **2003**, *13*, 211-219; b) Katsiari, M.C.; Voutsinas, L.P.; Alichanidis, E.; Roussis, I.G. *Int. Dairy J.* **2001**, *11*, 193-197; c) Georgala, A.; Moschopoulou, E.; Aktypis, A.; Massouras, T.; Zoidou, E.; Kandarikis, I.;Anifantakis, E. *Food Chem.* **2005**, *93*, 73-80.
[250] Fan, T.Y.; Hwang, D.H.; Kinsella, J.E. *J. Agric. Food Chem.* **1976**, *76*, 443-448.
[251] de la Fuente, M.A.; Fontecha, J.; Juárez, M. *Z. Lebensm. Unters. Forsch.* **1993**, *196*, 155-158.

Table 6.1. List of the FAs available from the standard mixtures, together with their experimental correction factors (CF) and retention times (RT(min)).

Analyzed compound	Original FA	Abbreviation	Molecular Formula	CF	RT(min)
methyl butyrate	butanoic acid	C4	$CH_3(CH_2)_2COOH$	1.38	2.61
methyl hexanoate (caproate)	caproic acid	C6	$CH_3(CH_2)_4COOH$	1.18	3.71
methyl octanoate (caprylate)	caprylic acid	C8	$CH_3(CH_2)_6COOH$	1.07	5.58
methyl decanoate (caprate)	capric acid	C10	$CH_3(CH_2)_8COOH$	1.02	7.69
methyl dodecanoate (laurate)	lauric acid	C12	$CH_3(CH_2)_{10}COOH$	0.97	9.70
methyl tridecanoate	tridecanoic acid	C13	$CH_3(CH_2)_{11}COOH$	0.96	10.62
methyl tetradecanoate (myristate)	myristic acid	C14	$CH_3(CH_2)_{12}COOH$	0.95	11.50
methyl *(9Z)*-tetradecenoate (myristoleate)	myristoleic acid	C14:1-*(9Z)*	$CH_3(CH_2)_3CH{=}CH(CH_2)_7COOH$	0.94	12.13
methyl pentadecanoate	pentadecanoic acid	C15	$CH_3(CH_2)_{13}COOH$	0.93	12.35
methyl hexadecanoate (palmitate)	palmitic acid	C16	$CH_3(CH_2)_{14}COOH$	0.93	13.29
methyl *(9Z)*-hexadecenoate (palmitoleate)	palmitoleic acid	C16:1-*(9Z)*	$CH_3(CH_2)_5CH{=}CH(CH_2)_7COOH$	0.92	13.91
methyl heptadecanoate (margarate)	margaric acid	C17	$CH_3(CH_2)_{15}COOH$	0.95	14.29
methyl octadecanoate (stearate)	stearic acid	C18	$CH_3(CH_2)_{16}COOH$	0.98	15.53
methyl *(9E)*-octadecenoate (elaidate)	elaidic acid	C18:1-*(9E)*	$CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$	0.96	16.06
methyl *(9Z)*-octadecenoate (oleate)	oleic acid	C18:1-*(9Z)*	$CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$	0.96	16.33
methyl *(9Z,12Z)*-octadecadienoate (linoleate)	linoleic acid	C18:2-*(9Z,12Z)*	$CH_3(CH_2)_4CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$	0.96	17.73
methyl icosanoate (arachidate)	arachidic acid	C20	$CH_3(CH_2)_{18}COOH$	1.00	18.81
methyl *(9Z,12Z,15Z)*-octadecatrienoate (alpha-linolenate)	alpha-linolenic acid	C18:3-*(9Z,12Z,15Z)*	$CH_3CH_2CH{=}CHCH_2CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$	0.96	19.77
methyl *(11Z)*-icosenoate (*cis*-11-eicosenoate)	*cis*-11-eicosenoic acid	C20:1-*(11Z)*	$CH_3(CH_2)_7CH{=}CH(CH_2)_9COOH$	1.07	19.95
methyl docosanoate (behenate)	behenic acid	C22	$CH_3(CH_2)_{20}COOH$	1.03	23.56
methyl *(13Z)*-docosenoate (erucate)	erucic acid	C22:1-*(13Z)*	$CH_3(CH_2)_7CH{=}CH(CH_2)_{11}COOH$	1.08	24.43
methyl tetracosanoate (lignocerate)	lignoceric acid	C24	$CH_3(CH_2)_{22}COOH$	1.06	26.52

Table 6.2. Relative abundance (mean and standard deviations, FA g/100g fat) of the identified FAs in PF, PR and PS. Significance (*p*-value) of difference in the means determined by one-way ANOVA and list of the significantly different class pairs according to Tukey-Kramer test.

	Triacylglycerols			Free fatty acids			ANOVA	Tukey-Kramer
Fatty acid	**PF**	**PR**	**PS**	**PF**	**PR**	**PS**	**(p-value)**	**(0.05 significance level)**
C6	3.8±0.5	3.8±0.5	4.4±0.7	2.5±0.5	4.8±0.5	3.3±0.7	$3.0*10^{-4}$	PS(TAG)-PF(FFA); PF(FFA)-PR(FFA); PR(FFA)-PF(FFA)
C8	3.2±0.4	3.4±0.7	3.7±0.5	2.3±0.4	4.0±0.7	3.9±0.5	$1.5*10^{-5}$	PR(TAG)-PF(FFA); PS(TAG)-PF(FFA); PF(FFA)-PR(FFA); PF(FFA)-PS(FFA)
C10	11±1	10±2	11±1	8±1	11±2	11±1	$8.6*10^{-4}$	**PF(TAG)-PF(FFA)**; PS(TAG)-PF(FFA); PF(FFA)-PR(FFA); PF(FFA)-PS(FFA)
C12	6±1	5±1	5.5±0.7	5±1	6±1	5.7±0.7	0.369	-
C14	14±1	13±1	13±1	13±1	12±1	11±1	$1.4*10^{-7}$	PF(TAG)-PR(FFA); PF(TAG)-PS(FFA); PR(TAG)-PS(FFA); **PS(TAG)-PS(FFA)**; PF(FFA)-PS(FFA)
C14:1-*(9Z)*	0.4±0.2	0.6±0.3	0.5±0.2	0.3±0.2	0.7±0.3	0.3±0.2	0.804	-
C15	1.4±0.2	1.2±0.3	1.1±0.5	1.2±0.2	1.1±0.3	1.0±0.5	0.030	PF(TAG)-PS(FFA)
C16	30±2	28±1	28±1	31±2	25±1	24±1	$6.0*10^{-13}$	PF(TAG)-PR(FFA); PF(TAG)-PS(FFA); PR(TAG)-PF(FFA); **PR(TAG)-PR(FFA)**; PR(TAG)-PS(FFA); PS(TAG)-PF(FFA); PS(TAG)-PR(FFA); **PS(TAG)-PS(FFA)**; PF(FFA)-PR(FFA); PF(FFA)-PS(FFA)
C16:1-*(9Z)*	0.7±0.2	0.5±0.2	0.4±0.2	0.9±0.2	0.7±0.2	0.9±0.2	$9.6*10^{-8}$	PR(TAG)-PF(FFA); PR(TAG)-PS(FFA); PS(TAG)-PF(FFA); PS(TAG)-PR(FFA); **PS(TAG)-PS(FFA)**
C17	0.7±0.1	0.6±0.2	0.5±0.2	0.7±0.1	0.5±0.2	0.6±0.2	0.008	PF(TAG)-PR(FFA); PF(FFA)-PR(FFA)
C18	10±3	12±1	12±1	10±3	9±1	10±1	$2.0*10^{-4}$	PF(TAG)-PR(TAG); **PR(TAG)-PR(FFA)**; PS(TAG)-PR(FFA)
C18:1-*(9E)*	0.9±0.2	0.9±0.3	0.9±0.4	1.2±0.2	2.1±0.3	1.7±0.4	$3.9*10^{-5}$	PR(FFA)-PF(FFA); **PR(FFA)-PR(TAG)**; PR(FFA)-PS(TAG)
C18:1-*(9Z)*	16±3	18±3	17±2	20±3	20±3	21±2	0.026	PF(TAG)-PS(FFA)
C18:2-*(9Z,12Z)*	2±1	1.9±0.6	1.8±0.7	3±1	2.8±0.6	3.4±0.7	$2.4*10^{-6}$	**PF(TAG)-PF(FFA)**; PF(TAG)-PS(FFA); PR(TAG)-PF(FFA); PR(TAG)-PS(FFA); PS(TAG)-PF(FFA); PS(TAG)-PR(FFA); **PS(TAG)-PS(FFA)**
C18:3-*(9Z,12Z,15Z)*	0.41±0.08	0.4±0.2	0.4±0.2	1.19±0.08	1.24±0.15	1.5±0.2	$3.6*10^{-13}$	**PF(TAG)-PF(FFA)**; PF(TAG)-PR(FFA); PF(TAG)-PS(FFA); PR(TAG)-PF(FFA); **PR(TAG)-PR(FFA)**; PR(TAG)-PS(FFA); PS(TAG)-PF(FFA); PS(TAG)-PR(FFA); **PS(TAG)-PS(FFA)**

Looking at Figure 6.2 and Table 6.2, a comparison of the percentage levels of the different fatty acids in the triglyceride and FFA fractions can be made. Significantly differing values for the medium-chain capric (C10, PF(TAG)-PF(FFA)), myristic (C14, PS(TAG)-PS(FFA)), palmitic (C16, PS(TAG)-PS(FFA) and PR(TAG)-PR(FFA)), palmitoleic acid (C16:1-*(9Z)*, PS(TAG)-PS(FFA)), stearic acid (C18, PR(TAG)-PR(FFA)), elaidic acid (C18:1-*(9E)*, PR(TAG)-PR(FFA)), linoleic (C18:2-*(9Z,12Z)*, PS(TAG)-PS(FFA) and PF(TAG)-PF(FFA)) and alpha-linolenic (C18:3-*(9Z,12Z,15Z)*, PS(TAG)-PS(FFA), PR(TAG)-PR(FFA) and PF(TAG)-PF(FFA)) acids can be observed. The levels of C16:1-*(9Z)*, C18:1-*(9E)*, C18:2-*(9Z,12Z)*, and C18:3-*(9Z,12Z,15Z)* were higher in the FFA fraction compared to the triglyceride fraction, while the saturated C10, C14, C16, C18 acids exhibit the opposite trend. This result is in-line with the indication that the manufacturing and ripening conditions are considerably significant for the formation of medium-chain length FFA in cheese[252].

6.2.2 Discrimination of Pecorino cheese samples

As above-mentioned, the aim of the present work is to develop a method of extraction of different constituents of cheese fat, in order to exploit TAG- and FFA-profiles to discriminate the three valuable Pecorino cheeses. In order to calibrate and validate the classification model, GC-data was split into two (training and test) representative sets of samples by the application of the Duplex algorithm on each class separately. Briefly, the TAG- and FFA- matrices of each class were subjected to ComDim[253] analysis; three common components (CCs) were calculated in each model and the three super-scores matrices ($\boldsymbol{T}_{PF}$, $\boldsymbol{T}_{PS}$ and $\boldsymbol{T}_{PR}$) obtained. Eventually, Duplex[235] algorithm was run individually on these latter ones and the samples were divided accordingly. Of the 50 Pecorino cheese samples, 29 (9 PF, 10 PR, and 10 PS) were selected as training set and 21 (6 PF, 8 PR, and 7 PS) employed for the validation of the model.

The performances of PLS-DA and MB approaches were compared both in calibration (through 5-fold cross-validation procedure) and in prediction. Table 6.3 reports the results in terms of correct classification rates (%) for each discriminant model calculated, together with the indication of the block(s) under investigation, the data normalization applied, and the number of latent variables (LVs) extracted (or of Original Variables (OVs) retained for SO-CovSel).

---

[252] Hayaloglu, A.A.; Karabulut, I. *Int. J. Food Prop*. **2013**, *16*, 1407-1416.
[253] Qannari, E.M.; Courcoux, P.; Vigneau, E. *Food Qual. Prefer*. **2001**, *12*, 365-368.

Table 6.3. PLS-DA and MB analysis. The performances of the models are reported in terms of correct classification rates (CCR,%), both in calibration (cross-validation) and external validation. In the MB-section, the pre-processing sequentially applied and the LVs (or OVs) extracted by the blocks are reported in square brackets.

Single-block Approach, CCRs(%)

				Training set (cross-validation)			External set (prediction)		
Model	**Block**	**Normalization**	**LVs**	**PF**	**PR**	**PS**	**PF**	**PR**	**PS**
PLS-DA	TAG	Autoscaling (A)	5	55.6	30.0	70.0	50.0	12.5	71.4
	TAG	mean-centering (MC)	2	55.6	10.0	80.0	50.0	25.0	57.1
	FFA	A	5	44.4	70.0	50.0	50.0	75.0	100.0
	FFA	MC	3	55.6	50.0	50.0	66.7	87.5	57.1

Multi-Block Approach, CCRs(%)

				Training set (cross-validation)			External set (prediction)		
Model	**Block**	**Normalization**	**LVs/OVs**	**PF**	**PR**	**PS**	**PF**	**PR**	**PS**
mid-level DF	TAG&FFA	A	2	66.7	60.0	60.0	66.7	62.5	100.0
		MC	3	66.7	60.0	70.0	66.7	75.0	57.1
SO-PLS-LDA	**TAG&FFA**	[A A]	[2 2]	77.8	80.0	80.0	50.0	50.0	57.0
		[MC MC]	**[5 2]**	**77.8**	**80.0**	**90.0**	**66.7**	**62.5**	**57.1**
SO-CovSel-LDA	TAG&FFA	[A A]	[4 3]	66.7	80.0	70.0	50.0	62.5	57.1
		[MC MC]	[0 1]	66.7	60.0	40.0	66.7	65.5	28.6

6.2.2.1 PLS-DA Analysis

As above-mentioned, the peak areas tables of TAG and FFA were divided into a calibration and a validation set. Autoscaling and mean-centering were tested on the training TAG- and FFA-profiles and compared. Four PLS-DA models were calculated, and the optimal number of latent variables (LVs) was extracted for each of them; the correct classification rates (CCRs%) obtained in a 5-fold cross-validation procedure are reported in Table 6.3. Inspecting it, it is possible to point out the inefficacy of the fingerprinting GC-FID profile of TAG and FFA fractions in the discrimination of the investigated Pecorino cheeses. The mean CCR for the single-block approaches, indeed, ranged from 48.5 to 54.8%, with the "best" performance achieved by handling the auto-scaled FA profile from free fraction alone.

6.2.2.2 MB Analysis

The classification was then attempted by Mid-Level data fusion, SO-PLS-LDA, and SO-CovSel-LDA, trying to improve the single-block approach performances taking advantage from the intrinsic characteristics, outlined in Section 3.4.3 of Chapter 3, of data fusion methods. In a Mid-Level framework, the traditional PLS-DA algorithm can be applied to the augmented matrix obtained by row-wise (sample-mode) concatenating the GC-profiles. In SO-PLS, the appropriate number of latent variables to keep is selected using the global optimization approach[167a] already described in Section 3.4.3 of Chapter 3. Briefly, cross-validated SO-PLS-LDA models were built with all the possible

combinations of LVs (within a fixed maximum of 10 for each block) and the optimal complexity was defined inspecting the cross-validated classification error rates. Also in this case, autoscaling and mean-centering were tested. Eventually, the optimal SO-PLS-LDA model was applied on the test set.

Inspecting Table 6.3, it can be highlighted that whatever MB approach is applied, a slightly improvement of CCRs is observed. SO-PLS-LDA, in particular, provides the better performances. In particular, mean CCR in cross-validation increases to 82.6% when the SO-PLS approach was applied on mean-centered TAG(5 LVs)- and FFA(2 LVs)-blocks. Nevertheless, when applying the model on the test set (normalized accordingly) the accuracy corresponded to 62.1% (corresponding to 8 misclassified test samples over 21; of these, 2 belong to PF, and 3 to PS and PR classes). The results of SO-PLS-LDA analysis are graphically shown in Figure 6.3, where the Pecorino samples are projected onto the space of the two canonical variates (CVs).

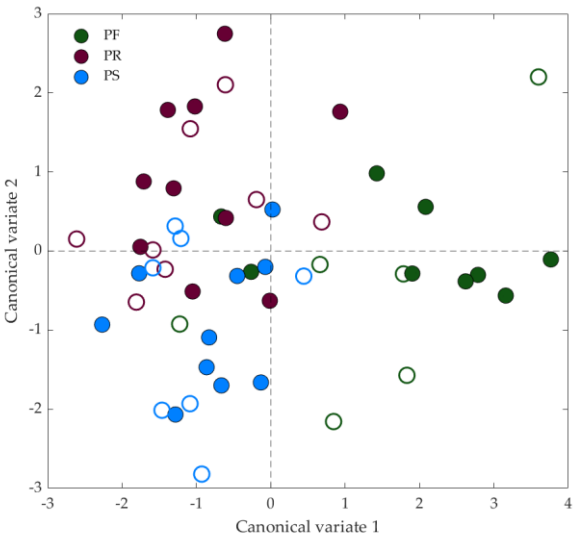


Figure 6.3. SO-PLS-LDA Analysis. Projection of the training (filled symbols) and test (empty symbols) samples onto the sub-space spanned by the two canonical variates (CVs).

Looking at this figure, it is possible to recognize a partial distinction of the individuals belonging to PF variety, and a large overlapping between the PR and PS samples. This observation, in line with the results pointed out in Table 6.3 for the class-CCRs under the single-block attempt, is consistent with the common origin of all the raw materials in the production of PR and PS PDO specialties and the dependency of fat composition from the provenance of milk, feeding, and so on (see Section 6.2.1). Moreover, two main

groupings can be recognized among PF samples distinguishing the cheeses produced in mechanized dairy plants from the ones made in small-scale cheesemaking processes generally carried out manually. This is in line with the findings provided in Section 5.2.2 of Chapter 5.

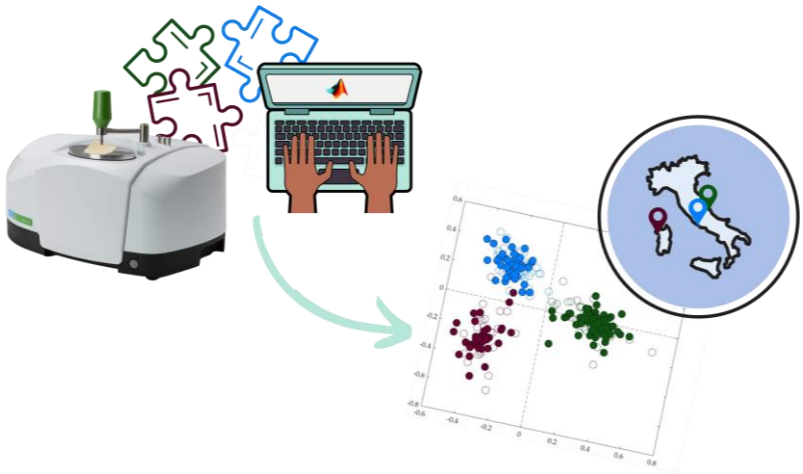
SO-CovSel-LDA carried out the classification selecting the caproic, the capric, the lauric, and the myristic (TAG fraction) acids and again the caproic, the lauric, and the pentadecanoic (FFA fraction) acids for autoscaled-data and only palmitic acid (FFA fraction) for the MC-profiles.

6.3 Conclusions

The aim of the present work was to build an analytical protocol for the characterization and discrimination of three typical and certified Pecorino cheeses produced in Italy with different and peculiar technologies: Pecorino di Farindola (Slow Food Presidium), Pecorino Sardo (PDO) and Pecorino Romano (PDO). For this purpose, the triacylglyceride (TAG) and free fatty acid (FFA) fractions of 50 Pecorino samples were analyzed by means of GC-FID. The data-blocks obtained were individually analyzed by PLS-DA or incorporated in multi-block models by using mid-level data fusion, SO-PLS or SO-CovSel. In general, all the data fusion approaches provided better results if compared with the one analyzing one profile at a time. Nevertheless, the classification performances were not fairly good to consider the fat composition a good marker for the geographical origin and the cheesemaking conditions.

Chapter 7

Application of SPORT algorithm on ATR-FTIR data: a rapid and green tool for the characterization and discrimination of three typical Italian Pecorino cheeses[68]



7.1 Materials and Methods

7.1.1 Pecorino cheese samples

Two hundred seventy-five (275) samples of the Pecorino cheeses (127 PF, 63 PR and 85 PS) were investigated.

More detailed information on the investigated Pecorino samples is reported in Section 4.1.1 of Chapter 4.

7.1.2 ATR FT-IR measurements

A PerkinElmer Spectrum Two™ (PerkinElmer, Waltham MA, USA) FT-IR spectrometer consisting in a deuterated triglycine sulfate (DTGS) detector and a PerkinElmer Universal Attenuated Total Reflectance (uATR) accessory equipped with a single bounce diamond crystal was used to record the infrared spectra from the slices of Pecorino

cheese. The background was collected with the crystal exposed to the air and each spectrum was registered from 4000 cm^{-1} to 400 cm^{-1} with 1 cm^{-1} instrumental resolution. Ten scans were averaged per spectral replicate. The force applied using the pressure monitoring system integrated in the instrument was tuned to maximize the spectrum intensity. The ATR crystal was cleaned with methanol and air-dried before each acquisition. The ATR FT-IR spectra were collected at different depths starting from the internal part of the slice and moving towards the cheese peel, following the naked eye-visible maturation gradient typical of the ripened cheeses.

7.1.3 Statistical Analysis

The ATR-FTIR signals collected as described in the Section 2.2 were exported from the Spectrum software (PerkinElmer, Waltham MA, USA) to the Matlab environment (R2020b; The Mathworks, Natick, MA). The imported data were converted to pseudo-absorbance (log(1/R)) to be further processed. To achieve the goal of classifying the samples according to their geographical origin and cheesemaking protocol, two classification methods were employed: PLS-DA, to handle the IR spectroscopic data with a unique pre-processing and testing each of them independently from one another, and SPORT, to achieve the extraction of complementary information from each of the differently pre-treated IR-blocks.

The Matlab functions employed in the present work can be freely downloaded from the RomeChemometrics website[221].

7.2 Results and Discussion

7.2.1 ATR-FTIR Characterization

In Figure 7.1 the ATR-FTIR average spectra collected on the PF, the PR and the PS samples are shown, reflecting the main constituents of cheese (proteins, fats, carbohydrates, and water).

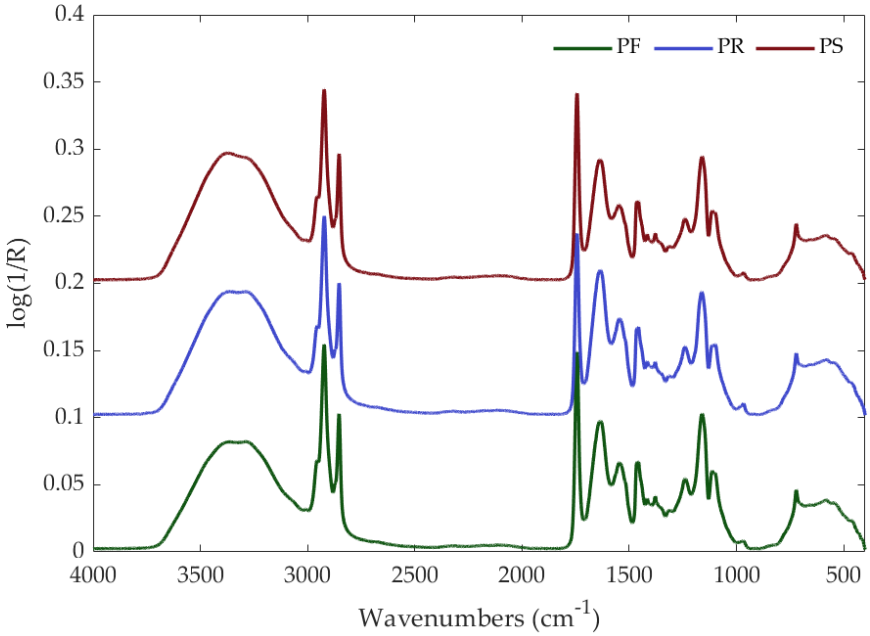


Figure 7.1. Mean ATR-FTIR spectra collected on PF, PR and PS samples.

EU official PDO reports[53] stipulated for PR and PS indicate only the fat content on dry matter (DM) as not less than 36%, 40% and 35% for PR, *dolce* PS and *maturo* PS, respectively. Further information about the composition of these Pecorino cheeses is not provided. PR has been investigated more in detail by Addis and collaborators[190] that reported 29-39g/100g of moisture, 43.4-50.9% of fat/DM, and 33-43% of protein/DM; DM is 68-70g/100g of cheese according to the study authored by Battistotti and Corradini[254]. PF is composed by 27% of proteins, 35% of fats, and 32% of moisture according to information provided by the official website of the related consortium[255].

Table 7.1 reports in details the hypothesis of assignment of the most important observed peaks, considering the main literature studies on this topic[256].

---

[254] Battistotti, B.; Corradini, C. In Fox, P.F. (Ed.) *Cheese: Chemistry, Physics and Microbiology*. *Volume 2: Major Cheese Groups*, 2nd edition (pp. 221-243). Boston, MA: Springer US, **1999**.

[255] Consorzio di Tutela del Pecorino di Farindola (italian source). Retrieved from https://www.pecorinodifarindola.it/disciplinare/ (last access January 2022).

[256] a) Ayvaz, H.; Mortas, M.; Dogan, M.A.; Atan, M.; Yildiz Tiryaki, G.; Karagul Yuceer, Y. *J. Food Sci. Technol*.**2021**, *58(10)*, 3981-3992; b) Bērziņš, K.; Harrison, S.D.L.; Leong, C.; Fraser-Miller, S.J.; Harper, M.J.; Diana, A.; Gibson, R.S.; Houghton, L.A.; Gordon, K.C. *Spectrochim. Acta-Part A Mol. Biomol. Spectrosc*. **2021**, *246*, 118982; c) Boubellouta, T.; Dufour, É. *Food Bioprocess Technol*. **2010**, *5*, 273-284; d) Coates, J. In Meyers, R.A. (Ed.)

Table 7.1. IR absorption frequencies identified in Pecorino cheese spectra.

Wavenumbers (cm^{-1})	Origin	Literature group frequency (cm^{-1})	Assignment	Representatives
3700-3015, peaks on 3386 and 3279	OH/NH	3570-3200 (broad)	hydroxy group, O-H stretch (H-bonded)/ N-H stretch	water, carbohydrates and proteins
3006	CH	3006	cis-olefinic groups	fats
2950	CH_3	2970-2950	terminal methyl C-H, asymmetric stretch	long-chain fatty acids
2920	CH_2	2935-2915	methylene C-H, asymmetric stretch	long-chain fatty acids
2870	CH_3	3000-2800	terminal methyl C-H, symmetric stretch	long-chain fatty acids
2853	CH_2	2865-2845	methylene C-H, symmetric stretch	long-chain fatty acids
1743	C=O	1750-1725; 1745	carbonyl group (ester), stretch	fatty acid esters and carbohydrates
1639	C=O/NH	1680-1630; 1630-1605	amide I (CO, stretch/ NH, wagging) and carboxylate (CO)	proteins and carbohydrates
1546	NH/CN	1550-1480	amide II (NH, bend/ CN, stretch)	proteins
1465	CH_3, C-C-H, C-O-H	1465; 1460-1340	methyl (asym.) CH, bend; =C–H *cis*-bending; C–C–H and C–O–H deformations	fats and carbohydrates (mostly lactic acid)
1456	CH, C-C-H, C-O-H	1453; 1460-1340	C–H bending vibrations and C–C–H and C–O–H deformations	fats and carbohydrates
1415	CH_2	1450-1410	various vibrational modes of CH_2 groups	fats, carbohydrates and proteins
1378	CH_3	1395-1350	methyl (sym.) CH, bend	fats and oils
1316	CH_2	1350-1150	various vibrational modes of CH_2 groups	fats, carbohydrates and proteins
1240	NH/CN, C-C/C-O	1330-1230; 1250-950	amide III (complex displacement); endo- and/or exocyclic C–C and C–O bonds	proteins, fats and carbohydrates
1160	C-O	~1175	ester linkage C-O	triacylglycerols (fats)
1114	C-O	1139-1105	CO, stretch	carbohydrates
1099	C-C/C-O	1250-950	endo- and/or exocyclic C–C and C–O bonds	carbohydrates
966	C=C	966	isolated out of plane *trans*-C∪C bending	fats

More in general, the following typical vibration patterns can be highlighted. The broad band centered at about 3386 and 3279 cm^{-1} can be assigned to the N-H stretching of proteins and O-H (H-bonded) stretching of carbohydrates and water. The two groups of

---

Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation (pp. 10815-10837). Hoboken, New Jersey, USA: John Wiley & Sons, **2006**; e) Fagan, C.C.; O'Donnell, C.P.; O'Callaghan, D.J.; Downey, G.; Sheehan, E.M.; Delahunty, C.M.; Everard, C.; Guinee, T.P.; Howard, V. *J. Food Sci.* **2007**, *72*, E130-E137; f) Karoui, R.; Mouazen, A.M.; Dufour, É.; Pillonel, L.; Picque, D.; Bosset, J.O.; De Baerdemaeker, J. *Le Lait*. **2006**, *86*, 83-97; g) Karoui, R.; Mouazen, A.M.; Dufour, É.; Pillonel, L.; Picque, D.; De Baerdemaeker, J.; Bosset, J.O. *Eur. Food Res. Technol*. **2005**, *222(1)*, 165-170; h) Mohamed, H.; Nagy, P.; Agbaba, J.; Kamal-Eldin, A. *Food Chem*. **2021**, *334*, 127436; i) Rodriguez-Saona, L.; Ayvaz, H.; Wehling, R.L. In Nielsen, S.S. (Ed.) *Food Analysis*. Part of the Food Science Text Series (pp. 107-127). Springer International Publishing, Cham, **2017**; l) Rodriguez-Saona, L.; Koca, N.; Harper, W.J.; Alvarez, V.B. *J. Dairy Sci*. **2006**, *89*, 1407-1412; m) Temizkan, R.; Can, A.; Dogan, M.A.; Mortas, M.; Ayvaz, H. *Int. Dairy J*. **2020**, *110*, 104795.

sharp signals at 2950 and 2870 cm^{-1} and 2920 and 2853 cm^{-1} are associated to antisymmetric and symmetric C-H stretching in terminal methyl groups and methylene units, respectively, being associated mainly to the contribution of fats.

The spectral region between 1770 and 1700 cm^{-1} takes origin from the combination of different vibrations involving fats: the first overtone of the C-H bonds and the C=O vibrations coming from R(CO)OH and R(CO)OR (detected peak at 1743 cm^{-1}, C=O stretch of fatty acid esters) structures. The signal at 1639 cm^{-1} can be assigned to the stretching of carbonyl of proteins (amide I) and to O-H (water) bending, while the signals at 1546 cm^{-1} and 1240 cm^{-1} can be attributed to the amide II and amide III bands (associated with coupled C–N stretching and N–H bending vibrations of the peptide group), with the additional contribution of endo- and/or exocyclic C–C and C–O bonds in the carbohydrates . The spectral region between 1474 and 1200 cm^{-1} takes origin from C-C-H, C-O-H, and O-C-H bending modes of carbohydrates, mainly lactic acid and lactose, in combination to asymmetric and symmetric bending of CH_3 in branched alkanes and =C–H cis-bending (fats). The spectral region between 1200 and 900 cm^{-1}, the so-called fingerprint region, is dominated by the C-O and C-C stretching modes of carbohydrates, by the C=O stretch in fats and by the C-O ester linkage vibrations in triacylglycerols.

7.2.2 Discrimination of Pecorino cheese samples

To pursue the external validation of the models, regardless of the classifier, the spectra were reorganized into a training set, for the optimization of the calibration model, and a test set, for validation. The class-matrices collecting the ATR-FTIR spectra differently pre-processed (raw data, SNV, 1st derivative, 2nd derivative) were subjected to ComDim[253] analysis; four common components (CCs) were calculated for each model and the three super-scores matrices ($\boldsymbol{T}_{PF}$, $\boldsymbol{T}_{PR}$ and $\boldsymbol{T}_{PS}$) obtained. Eventually, Duplex[235] algorithm was run individually on these latter ones and the spectra were divided accordingly. Of the 275 Pecorino cheese samples, 155 (76 PF, 35 PR and 44 PS) were selected as training set and 120 (51 PF, 28 PR and 41 PS) were employed as test set for the validation of the models.

The performances of PLS-DA and SPORT were compared in calibration (through 5-fold cross-validation) and only the most suitable approach (the one leading to the lowest classification error in cross-validation) was further tested in prediction. Table 7.2 reports the pre-processing approaches, the number of LVs extracted, and the cross-validated CCRs% of each discriminant model calculated on the training samples.

Table 7.2. Partial Least Squares-Discriminant Analysis (PLS-DA) and Sequential Pre-processing through ORThogonalization (SPORT) cross-validated correct classification rates (%) as function of pre-processing and complexity (LVs).

Method	Optimal number of LVs						Correct Classification Rate (%)		
	Raw	SNV	1st der.*	2nd der.§	SNV+1st der.*	SNV+2nd der.§	Training Set (**Cross-validation**)		
							PF	PR	PS
PLS-DA	8						85.53	94.29	88.64
		9					93.42	94.29	93.18
			10				**96.05**	**94.29**	**93.18**
				6			94.74	71.43	72.73
					8		90.79	91.43	93.18
						10	94.74	71.42	81.82
SPORT	1	2	9	0			**100.00**	**94.29**	**100.00**

* SG (Savitzky-Golay)-15-2-1
§ SG-15-3-2

7.2.2.1 PLS-DA Analysis

As above-mentioned, the ATR-FTIR signals were divided into a calibration and a validation set. Different pre-treatments were tested on the training spectra and compared with the results obtained from the raw data: Standard Normal Variate[257] (SNV, non-linear pre-treatment), 1st (linear pre-treatment) and 2nd derivative through the Savitzsky-Golay[258] filter (15 points window and a second or third order for the interpolating polynomial, respectively) and combinations of SNV and derivatives.

Even if not always explicitly mentioned, data is assumed to be mean-centered (MC) prior the creation of any model. A PLS-DA model was calculated, and the optimal number of latent variables (LVs) was extracted for each of the six pre-processing approaches; the correct classification rates (CCR%) obtained in a 5-fold cross-validation procedure are reported in Table 7.2. Inspection of this Table reveals that the models built on the data transformed by 2nd derivative (2nd der+MC and SNV+2nd der+MC) are the ones leading to the lowest mean CCRs (79.6 and 82.7%, respectively). Mean CCRs ranging from 89.5 to 94.5% were obtained with the other pre-treatments (and without pre-processing); finally, the best model was the one involving only the 1st derivative that provided a mean CCR of 94.5%.

[257] Barnes, R.J.; Dhanoa, M.S.; Lister, S.J. *Appl. Spectrosc.* **1989**, *43*, 772–777.
[258] Savitzky, A.; Golay, M.J.E. *Anal. Chem.* **1964**, *36*, 1627–1639.

7.2.2.2 SPORT Analysis

In SPORT, the same data-block associated with different pre-processing techniques are handled in a SO-PLS approach. The appropriate number of latent variables to keep for each pre-treatment was selected on the basis of the results of a cross-validation using the global optimization approach, as in Chapter 4. Briefly, cross-validated SO-PLS-LDA models were built with all the possible combinations of latent variables (within a fixed maximum of 12 for each pre-treatment) and the optimal complexity was defined inspecting the overall classification error rate. Eventually, a SO-PLS-LDA model using the suggested number of latent variables was built on the calibration set and applied on the test set.

Inspecting Table 7.2, it can be highlighted that the mean CCR in cross-validation increases to 98.1% when the SPORT approach was applied instead of the single-block models. The SPORT model gave the best performances when three out of four blocks were used in the multi-block process. In details, SO-PLS finally modelled the raw data (1 LV), the block pre-processed by SNV (2 LVs) and the predictor matrix pre-treated by first derivative (9LVs). The major contribution of the last pre-processing mode is consistent with the outcome of the PLS-DA approach, which showed the best performances when the 1st derivative was applied. The SPORT model applied on the test set (pre-processed accordingly) provided an accuracy of 98.3% (corresponding to 2 misclassified test samples over 120; of these, 1 belong to PF class and 1 to PR class). The results of SO-PLS-LDA analysis are graphically shown in Figure 7.2, where the Pecorino samples are projected onto the space of the two canonical variates (CVs).

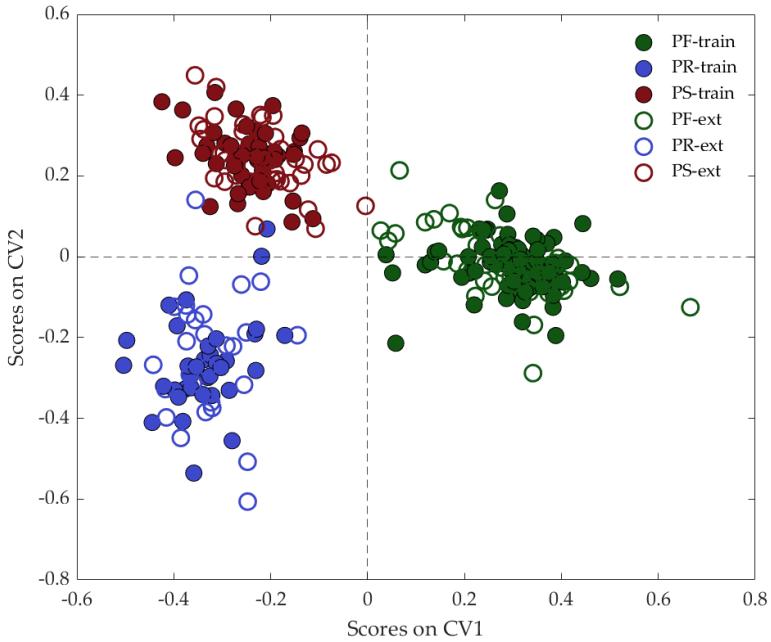


Figure 7.2. Sequential Pre-processing through ORThogonalization (SPORT) Analysis. Projection of the training (filled symbols) and test (empty symbols) samples onto the sub-space spanned by the first two canonical variate (CVs).

Looking at this Figure, it is possible to recognize a clear distinction among the individuals belonging to the three different cheese varieties. In particular, the objects of the PF class are clearly distinguishable along CV1 at positive scores whereas CV2 allows discriminating the samples belonging to PS class (second quadrant) from the PR category (third quadrant). In order to give a depth insight into the SPORT model, Variable Importance in Projection (VIP) indices were calculated following the embedded strategy described by Biancolillo and collaborators[259]. A graphical representation of VIP analysis, taking into account the most relevant block (1st derivative) is shown in Figure 7.3.

[259] Biancolillo, A.; Liland, K.H.; Måge, I.; Næs, T.; Bro, R. *Chemom. Intell. Lab. Syst.* **2016**, *156*, 89-101.

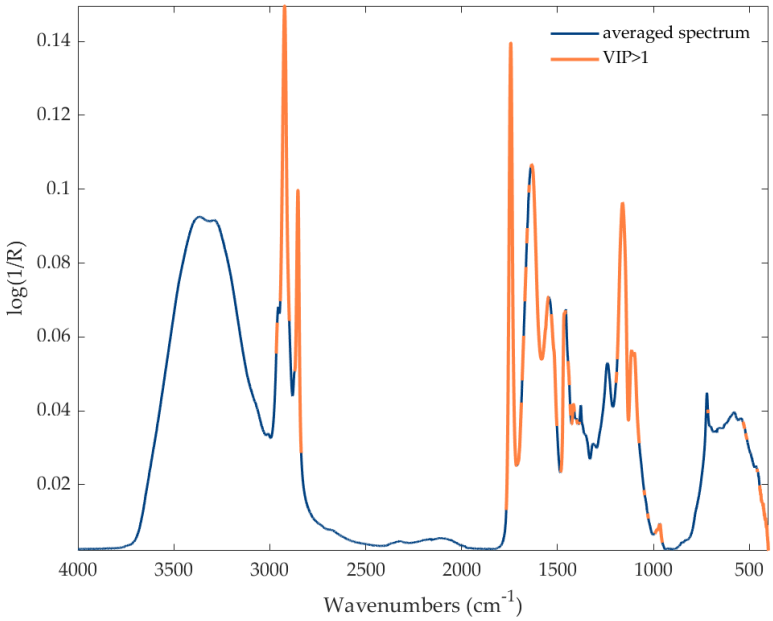


Figure 7.3. SPORT Analysis. Graphical representation of the 1st derivative-spectral variables presenting VIP index greater than 1. The average MIR spectrum is represented in blue, the variables with VIP > 1 are highlighted in orange.

Generally, each spectral variable presenting a VIP index higher than 1 (highlighted in orange) is considered to have a relevant contribution to the model. These are those from 2950 cm^{-1} to 2853 cm^{-1}, ascribable to the symmetric and antisymmetric stretching of the C-H bond in the long-chain fatty acids, from variables around 1743 cm^{-1} attributable to the stretching of carbonyl groups in the fatty acid esters and in the carbohydrates, those in the spectral range between 1639 cm^{-1} and 1378 cm^{-1} attributable to amide I and II modes, vibrations of carboxylate groups, and absorptions caused by different CH deformations in fats, proteins and carbohydrates, and those between 1160 cm^{-1} and 1099 cm^{-1} attributable to CO deformations. More details about the significant spectral absorptions can be found in Section 7.2.1 and in the related literature.

The possible effect of the order of the blocks on the SPORT performances was evaluated testing different combination orders of the four pre-treatments (Table 7.3).

Table 7.3. List of boosting of pre-treatments applied on the ATR-FTIR dataset, with different orders

Block number	Boosting 0	Boosting 1	Boosting 2	Boosting 3
1	raw data	SNV	SG-15-2-1	SG-15-3-2
2	SNV	raw data	SNV	SNV
3	SG-15-2-1	SG-15-3-2	raw data	SG-15-2-1
4	SG-15-3-2	SG-15-2-1	SG-15-3-2	raw data
#LVs	1,2,9,0	0,2,0,10	10,0,0,4	0,2,10,1
mean c.v.-CCR (%)	98.10	99.05	100.00	99.05
mean prediction-CCR (%)	98.16	94.96	93.34	94.31

It can be noted that only a slight difference, both in cross-validation and in prediction, results from changing the order of the blocks. The total number of latent variables varies from 12 to 14, and it is distributed differently over the selected blocks. Nevertheless, 1st derivative is the only one always selected and 2nd derivative is selected only once, in line with the previous outcomes.

7.3 Conclusions

This investigation has explored the possibility of developing a green, rapid and non-destructive approach suitable for the discrimination of typical Pecorino cheeses. In order to achieve this goal, samples were analyzed by ATR-FTIR Spectroscopy and handled using a novel approach to associate and select pre-treatments of spectra before their use in a calibration model with classification purpose. Indeed, the data-blocks obtained were individually analyzed by PLS-DA or incorporated in multi-block models by an approach called Sequential Pre-processing through ORThogonalization (SPORT), belonging to boosting family of ensemble methods. In general, both approaches provided good results from the prediction point of view. Concerning the classification performance pursued on the individual data-blocks, the lowest classification error (94.5% accuracy in cross-validation) was provided by the PLS-DA model calculated on first-derivative spectra. Nevertheless, the best results were provided by the data fusion strategy which allowed achieving extremely satisfactory results, misclassifying only two out of the 120 samples constituting the external validation set (98.3% accuracy).

Chapter 8

Sequential Data Fusion technique for the characterization and discrimination of typical Pecorino cheese produced in Central Italy

8.1 Materials and Methods

8.1.1 Pecorino cheese samples

Data collected on pecorino cheeses exploited in the previous chapters were taken under consideration in a multi-block framework; consequently, only samples analyzed by the different platforms were examined, circumscribing MB analysis to a common subset of objects among the different techniques. The volatile profile data set discussed in Chapter 4 presents the most reduced set of samples, and, therefore, it was the limiting-block for this further investigation.

Eventually, forty-seven (47) samples of PF (14), PS (16), and PR (17) were considered; further information can be retrieved from Section 4.1.1 of Chapter 4.

8.1.2 Statistical Analysis

The following predictor $\boldsymbol{X}$-blocks were considered: i) the peak areas of 39 identified volatile organic compounds (VOCs), previously subjected to logarithmic transformation; ii) the concentration values (Ba, Ca, Fe, K, Mg, Na, P and Zn μg/g of dried sample) of the multi-element composition (M-EC); iii) the peak areas (corrected for CF) of the 15 identified FAs in the TAG fraction of total fat; iv) the peak areas (corrected for CF) of the 15 identified FAs in the FFA fraction of total fat; v) the ATR-FTIR signals. PLS-DA, carried out as single-block benchmark method, was compared with SO-PLS-LDA, with the aim of reaching enhanced performances of the origin/cheesemaking process-classification problem outlined in this Section of the thesis.

The Matlab functions employed in the present work can be freely downloaded from the RomeChemometrics website[221].

8.2 Results and Discussion

8.2.1 Discrimination of Pecorino cheese samples

Data splitting was performed on the super-scores matrices ($\boldsymbol{T}_{PF}$, $\boldsymbol{T}_{PR}$ and $\boldsymbol{T}_{PS}$) obtained by applying the ComDim[260] analysis on the five $\boldsymbol{X}$-blocks, individually for each class. Five common components (CCs) were calculated for each model. Eventually, Duplex[235] algorithm was run individually on these latter ones to split the samples. Of the 47 Pecorino cheese samples, 27 (8 PF, 9 PR, and 10 PS) were selected as training set and 20 (6 PF, 8 PR, and 6 PS) employed for the validation of the model. Therefore, the reader should consider that the calibration and prediction subsets are different from the ones employed in each of the previous chapters; the samples are assigned to one set or another basing on all the collected predictor blocks.

The performances of PLS-DA and SO-PLS-LDA approaches were compared both in calibration (through 5-fold cross-validation procedure) and in prediction. Table 8.1 reports the results in terms of correct classification rates (%) for each discriminant model calculated, together with the indication of the block(s) under investigation, the data pre-processing applied, and the number of LVs extracted.

---

[260] Qannari, E.M.; Courcoux, P.; Vigneau, E. *Food Qual. Prefer*. **2001**, *12*, 365–368

Table 8.1. PLS-DA and MB analysis. The performances of the models are reported in terms of mean CCR,%, both in calibration and validation.

Single-block Approach, Mean CCR(%)

Model	Block	Pre-processing	LVs	Training set (cross-validation)	External set (prediction)
PLS-DA	VOCs (Chapter 4)	log_{10}+A	3	95.83	90.28
	M-EC (Chapter 5)	**A**	**6**	**96.67**	**95.83**
	TAGs (Chapter 6)	MC	13	89.63	44.44
	FFAs (Chapter 6)	MC	13	88.80	59.72
	ATR-FTIR (Chapter 7)	SNV+MC	15	100	88.89
		1st der.+MC	**7**	**100**	**90.28**
		SNV+1st der.+MC	8	100	84.72

Multi-block Approach, Mean CCR(%)

Model	Block	Pre-processing	LVs	Training set (cross-validation)	External set (prediction)
SO-PLS-LDA	**M-EC**	**A**	**3**	**100**	**95.83**
	TAGs	**MC**	**0**		
	FFAs	**MC**	**0**		
	VOCs	**log_{10}+A**	**2**		
	M-EC	A	0	100	94.44
	TAGs	MC	0		
	FFAs	MC	0		
	VOCs	log_{10}+A	4		
	ATR-FTIR	1st der+MC	5		

Abbreviations: A, autoscaling; MC: mean-centering; SNV: Standard Normal Variate; 1st der, Savitzky-Golay filter-15-2-1

Two different attempts were tested in a multi-block framework, following two trends currently discussed in literature, i.e., the possibility of carrying out a multi-set characterization by means of several platforms, and, on the other hand, the opportunity of accomplishing the same aim, but exploiting only *untargeted* techniques.

From one side, due to the chemical complexity of the food matrices and thanks to the extraordinary effort of multi-block approaches, a multi-methodological characterization employing different analytical platforms is nowadays often performed, usually leading to an improvement of the model performances. For this reason, the results of the SO-PLS-LDA method applied on all the five blocks together are here reported. From here on, this approach will be referred to as *targeted&untargeted*.

On the other side, food-authorities are asking for an extended list of fast, green, cheap, and easily applicable (at a custom level and *inline* in the manufactures) analytical methods. In the majority of the cases, like in this thesis, all these attributes are fulfilled by using *fingerprinting untargeted* techniques.

Could they completely replace the *targeted* approaches? *Targeted* techniques are surely more resource-intensive, but they are capable to provide huge amounts of information, often fundamental in traceability problems. To answer to this question, the results of SO-

PLS-LDA model developed only on the *targeted/convoluted* techniques (VOCs, M-EC, TAGs, and FFAs), hereinafter referred to as *only-targeted*, are here compared with the ones obtained using the *fingerprinting* ATR-FTIR-profiles, hereinafter referred to as *only-untargeted*.

8.2.1.1 PLS-DA Analysis

During the calibration of the different single-block expedients, different pre-treatments were tested on the training set and compared with the results (in cross-validation) obtained from the raw data: SNV, 1st derivative through the Savitzsky-Golay filter (15 points window and a second order for the interpolating polynomial) and a combination of SNV and derivative. Data were always auto-scaled or mean-centered prior to the creation of any model. In Table 8.1 only the best approaches are reported. A PLS-DA model was calculated, and the optimal number of latent variables (LVs) was extracted for each of the five data-blocks; the correct classification rates (CCR%) obtained in a 5-fold cross-validation procedure are reported in Table 8.1.

Inspection of this Table reveals that the models built on the data from the *targeted* ICP-OES and the *untargeted* (transformed by 1st derivative+MC) ATR-FTIR techniques are the ones leading to the higher mean CCRs (96.7 (c.v.)-95.8 (pred.) and 100.0(c.v.)-90.3% (pred.), respectively). Mean CCRs ranging from 88.8 to 95.8% (in cross-validation) were obtained in the other cases; the weaker models, as expected, were the ones involving the FAs profiles.

All the models, with the exception of the one built on ATR-FTIR data (this point will be discussed below), present better performances in prediction if compared with the single approaches reported in the respective chapters. This result could be due to the positive effect of a complete characterization of the samples on the representativeness of the sets generated by the Duplex algorithm in the data splitting procedure.

8.2.1.2 SO-PLS-LDA Analysis

In the two SO-PLS models here developed, the appropriate number of latent variables to keep for each block was selected on the basis of the results of a cross-validation using the global optimization approach briefly described in Chapters 4 and 6. Eventually, a SO-PLS-LDA model using the suggested number of latent variables was built on the calibration set and applied on the test set.

First of all, inspecting Table 8.1, it can be highlighted that the mean CCRs in both cross-validation and prediction reach their maximum when the MB approach, regardless of the blocks involved in the data fusion (i.e., *targeted&untargeted* vs *only-targeted*), was applied instead of the single-block models.

When ATR-FTIR data set was excluded from the handling, i.e., in the *only-targeted* approach, SO-PLS retained two out of four blocks in the multi-block process. More in details, only M-EC (3 LVs) and VOCs (2 LVs), the ones better performing even in the previous chapters, were used as predictor matrices. However, when ATR-FTIR data was included (i.e., *targeted&untargeted* case) in the data fusion, SO-PLS retained it together with VOCs. Due to the intrinsic characteristics of sequential methods, the blocks-selection varies when different data blocks are provided to the algorithm. ATR-FTIR data, when available, is selected.

The SO-PLS models applied on the test set (pre-processed accordingly) provided a mean CCR of 94.4% (corresponding to 1 PS misclassified test sample over 20) in the *targeted&untargeted-approach*, and of 95.8% (corresponding to 1 PR misclassified test sample over 20) in the *only-targeted* one. This latter can be compared with the performances of the *only-untargeted* model, providing a mean CCR of 90.3% (1 PR and 1 PS misclassified samples in prediction).

The projection onto the space of the two canonical variates is graphically shown in Figure 8.1.

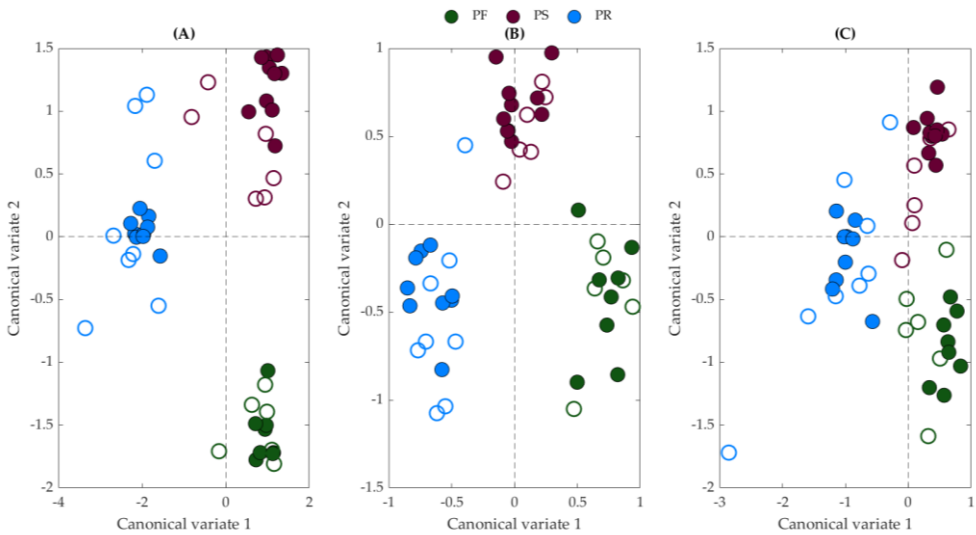


Figure 8.1. SO-PLS-LDA Analysis. Projection of the training (filled symbols) and test (empty symbols) samples onto the sub-space spanned by the first two canonical variate (CVs). (A) *targeted&untargeted*; (B) *only-targeted*; (C) *only-untargeted*.

The *only-targeted* approach built on a limited number of samples provides weaker performances in predicting the origin of the Pecorino cheeses, if compared with the others single approaches (M-EC and VOCs), and both the MB strategies. Among them,

better performances are reached when ATR-FTIR data is excluded; nevertheless, it is selected in the SO-procedure when available, meaning that relevant additional information is however provided by this *untargeted* technique.

8.3 Conclusions

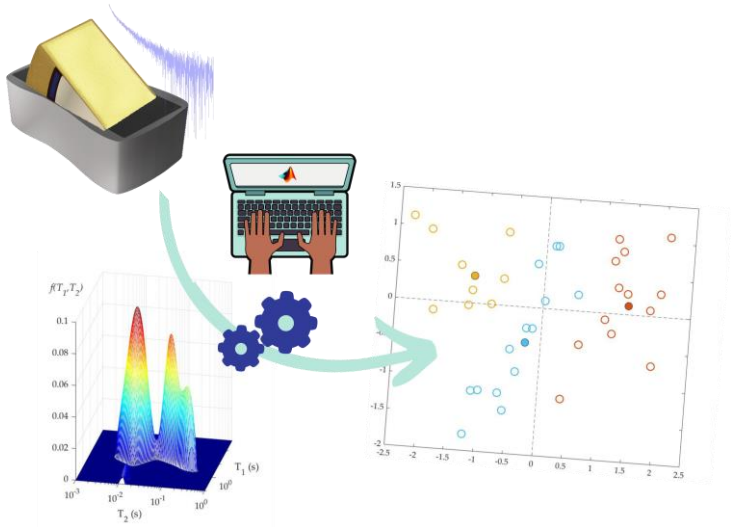
This Chapter has explored the possibilities and implications in the selection of an experimental set-up/approach among all the possible ones. *Untargeted* techniques are nowadays an invaluable resource in food traceability and authentication making possible fast, low-cost, and easy-to-handle applications. Nevertheless, a large data set of training samples, allowing to develop reliable and robust predictive models, is often required when dealing with this kind of techniques. This is the case of the use of ATR-FTIR in the discrimination of the three high-valued Italian PF, PR and PS described in Chapter 7, where the highest mean CCR in prediction (98.2%) among all the attempts carried out in Section 1 is reached, even better than its merge with the outcome of *targeted* platforms. This is a real challenge for the current times: we have the instrumentations and the skills to perform economic- and environmentally-compliant daily controls in the food supply-chain (avoiding the random controls using the *convoluted* technologies), but any effort in this direction will be meaningless without the institution by the competent authorities of intensively campaigns of creation and updating of largely food products-data sets.

Section 2

Green approaches for the characterization *inline* of dairy commodities

Chapter 9

Portable 2D-NMR-R technique for the storage monitoring of Pecorino cheeses: a novel approach for acquisition parameters optimization using Chemometrics



This project was carried out in collaboration with Prof. Nicoletta Spreti and Prof. Cinzia Casieri's research group (University of L'Aquila) and is currently not disclosed.

9.1 Introduction

In food processing, it is important to characterize the raw materials and the final products, e.g., moisture, fat, and sugar contents, of the continuous stream involved in the manufacturing process. Even more, it would be important to perform in real-time non-invasive measurements of the material stream for process and quality control purposes. During the chemical processes involved in the making and storage of a food commodity, indeed, many transformations, at different stages of the manufacturing system, take place[261]. Dairy systems, in particular, are fascinating for the scientists because they offer a large collection of food products that need to be characterized with designed applications. Dairy products can be considered as colloidal systems of fat globules in a

[261] Song, Y.-Q. *Prog. Nucl. Magn. Reson. Spectrosc.* **2009**, *55*, 324-334.

continuous phase constituted by water. Since the diffusion of fat and water molecules are affected by the presence of such suspension, this microstructure could be quantified by relaxation and diffusion measurements. NMR relaxation indeed, is a flexible, relatively inexpensive (if compared with high-resolution NMR), and customizable technique potentially suitable for *inline* analysis; nevertheless, and surprisingly, it is still considered a niche technique in food analysis. Moreover, it is usually not used at its fullest possibilities. Indeed, the investigation of the effects of pH on the casein gel structure[262], of the addition of emulsifying salt[263], of heating[264], and of the different processing stages in the manufacture of imitation cheese[265] has been proposed in literature by looking only to T_2 evolution, or to both T_1 and T_2 decays, but separately, i.e., using the 1D NMR-R.

The researchers Song and Hürlimann have been the leading pioneer in the application of 2D NMR-R to food problems, with a particular interest towards dairy products. Nevertheless, Chemometrics is totally excluded by the state of art of this kind of applications.

In this work, ANOVA-Simultaneous Component Analysis has been applied to monitor the evolution over three years of 2D-NMR-R profiles of fourteen Pecorino cheese samples with different ripening times (i.e., starting moisture content) from Central Italy. This preliminary study investigates indeed the potentiality of the application of such portable NMR-R probes in the routinely quality control in food supply-chain.

Moreover, the active research topic[266] of the choice of the most suitable regularization parameter α, being a trade-off between avoiding bias and insuring stability, is also addressed in this work. It is a key factor in the control of the smoothness of the acquired signals involved in the regularization methods for the solution of the ill-conditioning of Laplace inversion. A novel approach finalized to its definition is here successfully proposed; the correlation distribution functions have been handled at this purpose by the multivariate explorative approach of Principal Component Analysis, exploiting its diagnostic nature.

---

[262] a) Andersen, C.M.; Frost, M.B.; Viereck, N. *Int Dairy J*. **2010**, *20(1)*, 32-39; b) Moller, S.M.; Hansen, T.B.; Andersen, U.; Lillevang, S.K.; Rasmussen, A.; Bertram, H.C. *J Agric Food Chem*. **2012**, *60(7)*, 1635-1644.
[263] Chen, L.; Liu, H. *J Dairy Sci*. **2012**, *95(9)*, 4823-4830.
[264] Vogt, S.J.; Smith, J.R.; Seymour, J.D.; Carr, A.J.; Golding, M.D.; Codd, S.L. *J Food Eng*. **2015**, *150*, 35–43.
[265] Noronha, N.; Duggan, E.; Ziegler, G.R.; O'Riordan, E.D.; O'Sullivan, M. *Food Hydrocoll*. **2008**, *22(8)*, 1612-1621.
[266] Hürlimann, M.D.; Burcaw, L.; Song, Y.Q. *Journal of Colloid and Interface Science*. **2006**, *297*, 303-311.

9.2 Materials and Methods

9.2.1 Pecorino cheese samples

This preliminary study was carried out on 14 Pecorino cheeses with different degree of ripening produced in Central Italy (Abruzzo and neighboring areas) in 2019. Samples were collected according to their labeling in commerce and their moisture content, calculated by means of a 24-hours drying using the drying oven Büchi T0-51. A set of 5 Pecorino cheeses with a high moisture content (humidity higher than 30%, hereinafter referred to as *hm* (high-moisture)), 5 cheeses with a mid-moisture content (humidity between 22 and 30%, hereinafter referred to as *mm*), and 4 characterized by a low content in water (less than 22%, hereinafter referred to as *lm*) were eventually available. The cheese samples were packaged under-vacuum and analyzed not later than 10 days from the purchase, then stored in a refrigerator at 4°C. The analysis of the three different groups of cheeses was performed again after one (2020), and two years (2021).

9.2.2 2D ^{1}H NMR-R measurements

The two-dimensional proton nuclear magnetic resonance relaxometry (2D ^{1}H NMR-R) experiments were carried out using the mq-ProFiler (Bruker Biospin, Italy) as NMR equipment at low magnetic field intensity. It consists of a single-sided surface probe for ^{1}H nuclei and a portable electronic apparatus. The coil in use works at a Larmor frequency of 17.8 MHz, which can excite the sample up to a depth of about 2 mm and with a sensitive volume (x, y, z) of about 2×2×0.8 cm^{3}. An entire vacuum-packed slice of cheese was placed on the probe, making the measurements non-destructive. The experimental set-up is schematized in Figure 9.1.

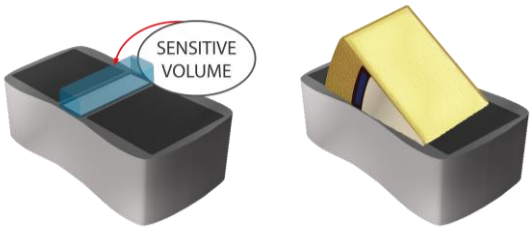


Figure 9.1. Scheme of measurement arrangement for NMR relaxation acquisitions.

The correlation between the longitudinal (T_1) and transverse relaxation (T_2) times was collected using a radiofrequency pulse sequence consisting of two editing parts, i.e., the evolution and the detection period, during which the spin system evolves respectively under the T_1 and T_2 relaxation mechanisms. More in details, relaxation data were gained through a saturation recovery and a CPMG pulse sequence described as $[saturation - \tau_S - \pi/2 - \tau_E(\pi - \tau_E - acquisition - \tau_E -)_{nE} - \tau_{RD}]_m$, where $\pi/2$ and π indicate the standard radio-frequency pulses, and saturation represents ten $\pi/2$ pulses of fixed length. In the acquisition of T_1-T_2 correlated signals, the T_2 encoding time, i.e., τ_E, was established for different T1 encoding times, i.e., τ_S. With the aim of detecting a large interval of relaxation times, the total T2 encoding time ($2\tau_E n_E$) was set to ~600ms of transverse relaxation decay with $2\tau_E = 300\mu s$ and $n_E = 2000$, whereas the T1 encoding was performed ranging τ_S from 1ms to 10s, according to a geometric progression with $n_S = 51$ different values. Each signal acquisition was averaged over $m = 62$ scans with $\tau_{RD} = 2s$ as recycle delay time for improving the signal-to-noise ratio. The $n_E x n_S$ two-dimendional data was processed using the 2D Fast Laplace inversion algorithm proposed by Song and collaborators[111] in order to obtain the T_1-T2 relaxation distribution function $f(T_1, T_2)$. The optimal regularization parameter α was set through a multivariate approach described in Section 9.3.2.

9.2.3 Statistical Analysis

The NMR experiments are designed as follow: Pecorino samples, being characterized by three different ripening times (factor β: *Maturation*, 3 levels: high-, mid-, and low-moisture content), have been analyzed at three different times (factor α: *Years*, 3 levels: 2019, 2020, and 2021). The ANOVA-like decomposition (eq. 3.21) was applied on the NMR data:

$$\|\boldsymbol{X}_c\|^2 = \|\boldsymbol{X}_{Years}\|^2 + \|\boldsymbol{X}_{Maturation}\|^2 + \|\boldsymbol{X}_{Years*Maturation}\|^2 + \|\boldsymbol{X}_e\|^2 \qquad \text{(eq. 9.2)}$$

The different contributions in the case of the data set considered in this study are the term accounting for the effect of time ($\boldsymbol{X}_{Years}$), the one accounting for the initial state of the sample ($\boldsymbol{X}_{Maturation}$) and the residual variation ($\boldsymbol{X}_e$), i.e., the random variation among the samples not accounted by the model. To assess the significance of the two main effects, the procedure based on permutation test (using 10000 randomizations) was adopted, as described in Section 3.3.3 of Chapter 3. In the SC Analysis, systematic information from the effect matrix were derived. Two components were extracted for each factor, both studied at three levels. To simultaneously evaluate the systematic variation among the averages of the measurements at the same levels and the natural (randomic) variability within each factor's level, the modified score plot derived recently

proposed by Zwanenburg and collaborators[267] was built, i.e., the residual term ($\boldsymbol{X}_e$) is projected onto the principal component subspace spanned by the effect matrix.

The theory of the mentioned Chemometric techniques is described in Chapter 3.

The Matlab functions employed in the present work can be freely downloaded from the RomeChemometrics website[221].

9.3 Results and Discussion

9.3.1 2D ^{1}H-NMR-R measurements and processing

Figure 9.2 shows the measured T_1-T_2 correlation map, identifying different chemical populations depending on (T_1,T_2) coordinates, of one selected cheese with high moisture level.

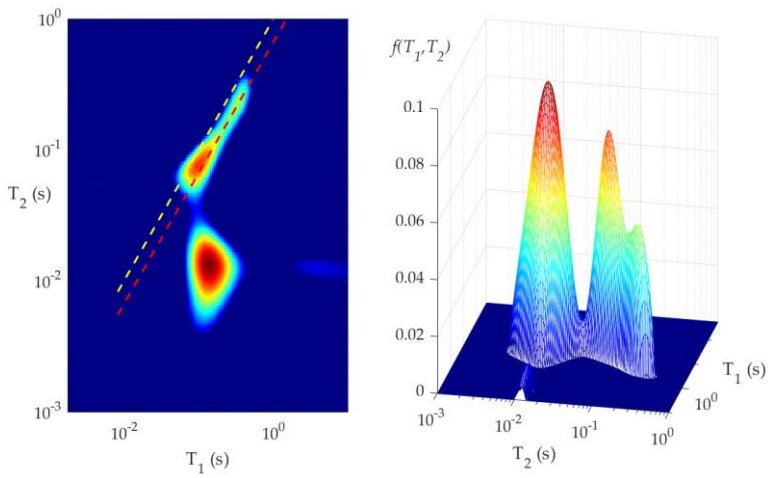


Figure 9.2. Two-dimensional nuclear magnetic resonance map. The dashed yellow line corresponds to T_1=T_2, whereas the red one indicates the quasi-diamagnetic reference line T_1=1.5T_2.

The signals from two different chemical components, although not fully separated, are present in the T_1-T_2 distribution. Briefly, dairy products are a mixture of water and fat, with some macromolecules (i.e., casein proteins, and lactose) localized at the water-fat interface. Therefore, the bimodal distribution may arise from the superposition of a

[267] Zwanenburg, G.; Hoefsloot, H.C.J.; Westerhuis, J.A.; Jansen, J.J.; Smilde, A.K. *J. Chemom*. **2011**, *25*, 561–567.

relatively sharp peak with a large T_1/T_2 ratio associated with the aqueous phase, and a broader feature characterized by a wider T_2 distribution due to the liquid fat content having a T_1/T_2 ratio close to 1. The solid fat and protein signals decay too quickly for our measurements, so would be not detected here. The observed relaxation times associated with water ($T_1/T_2>1$) are significantly lower than the bulk-water values ($T_1=1.5T_2$, dashed red line in Figure 9.1), indicating that the detected water molecules are in fast exchange between those internal to the casein continuous porous network and labile protein protons, and free water. Moreover, T_2 relaxation results to be faster than T_1 relaxation, indicating that exchange of water molecules or protons occurs at a rate comparable to (or slower than) the Larmor frequency. The fat composition in cheese, almost exclusively influenced by the fat profile in milk (as assessed in Chapter 6), includes molecules (predominantly TAGs) with a wide range of aliphatic chain lengths (the size of the three FAs in TAG can differ significantly). This distribution of molecular size results in non-exponential relaxation decays[268], consistently reflected in the wide T_2 distribution of the fat signal. The assignation of features here performed can be further supported looking at the Figure 9.3, showing the modifications occurring in the correlation map with the increasing of the ripening time of the cheeses.

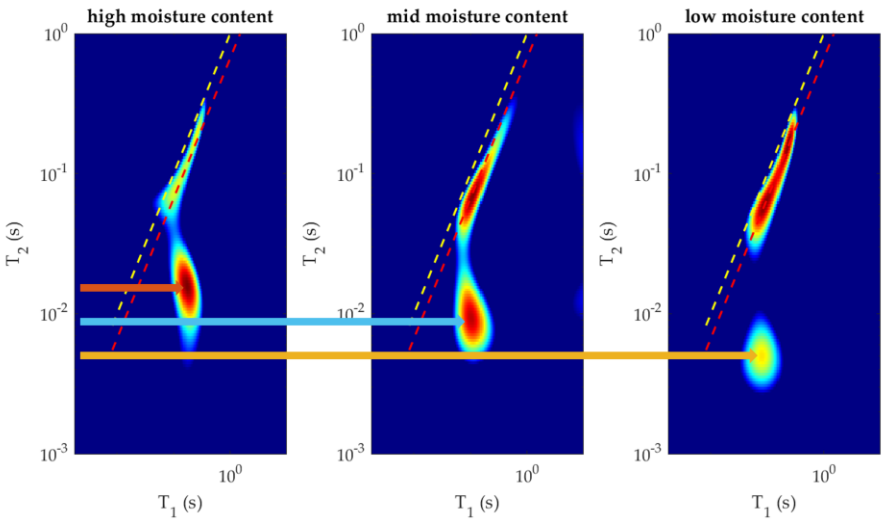


Figure 9.3. Comparation of relaxation maps from cheeses with different content in moisture. The dashed yellow line corresponds to $T_1=T_2$, whereas the red one indicates the quasi-diamagnetic reference line $T_1=1.5T_2$.

[268] Métais, A.; Mariette, F. *J. Magn. Reson.* **2003**, *165(2)*, 265-275.

The cheese with a longer time of maturation presents typically an higher concentrations of its chemical constituents (mainly fats, casein globules, and salts), due to a gradual decreasing in water content. Because T_2 relaxation time for water protons is mainly controlled by the concentration of casein[269], we can observe a more pronounced reduction of relaxation time (the water signal moves to shorter relaxation times) with increasing of ripening time. The intensity of the distribution functions related to the content in liquid fat seems to decrease, accordingly, but its position in the T_1-T_2 plane is unaffected. Moreover, the higher resolution observed between the two domains for the low moisture content cheese can also be justified by a lower extent of the protons exchange.

9.3.2 Selection of the smoothing coefficient α

Figure 9.4 shows the dependence of the extracted distribution functions $f(T_1,T_2)$ on the regularization parameter α, going from the minimum to the maximum α value encountered in the analysis of cheese samples.

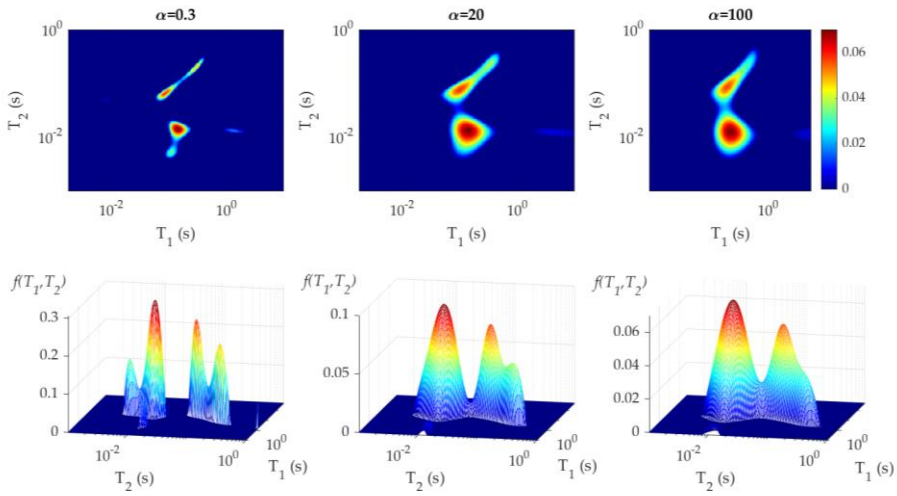


Figure 9.4. $f(T_1,T_2)$ extracted for the same *hm*-Pecorino cheese using three different values of α.

[269] a) Le Dean, A.; Mariette, F.; Marin, M. *J. Agric. Food Chem*. **2004**, *52(17)*, 5449-5455; b) Gottwald, A.; Creamer, L.K.; Hubbard, P.L.; Callaghan, P.T. *J. Chem.Phys*. **2005**, *122*, 034506; c) Mariette, F. In Webb, G. (Ed.) *Modern Magnatic Resonance*, 2nd edition (pp. 1535-1557). Springer, Cham, **2018**.

The two main features previously described are present in all cases; nevertheless, additional spurious signals appear for small values of α and a significant broadening occurs for larger values of the smoothing parameter.

With the aim of handling the 2D ^{1}H NMR-R data using the explorative multivariate ASCA approach, the influence of α on the projection of samples in the principal components space is additionally investigated. Principal component analysis was performed once a reasonable number of samples had been acquired. The data matrix $\boldsymbol{X}$ (34x22500) collecting the unfolded (22500=150 (T_1 dimension) *150 (T_2 dimension)) $f(T_1,T_2)$ distributions of each cheese was available. The $\boldsymbol{X}$ matrix was mean-centered before the analysis. The selected *lm*/*mm*/*hm* cheese acquired at different α values (0.3, 5, 10, 15, 20, 25, 30, 60, and 100) was back-projected on the model calculated on the 25 remaining (acquired up to that point) samples at their own experimental α-values. Hotelling T^2 and Q residuals statistic was applied to evaluate the behavior of samples at different α in the calculated models, whose optimal complexity was established using the Scree plot criteria. Figure 9.5 shows the results obtained for three selected Pecorino cheeses, one for each level of ripening time; whereas Figure 9.6 highlights the trend of (PC1, PC2) coordinates for all the acquired samples in 2020 as function of α parameter.

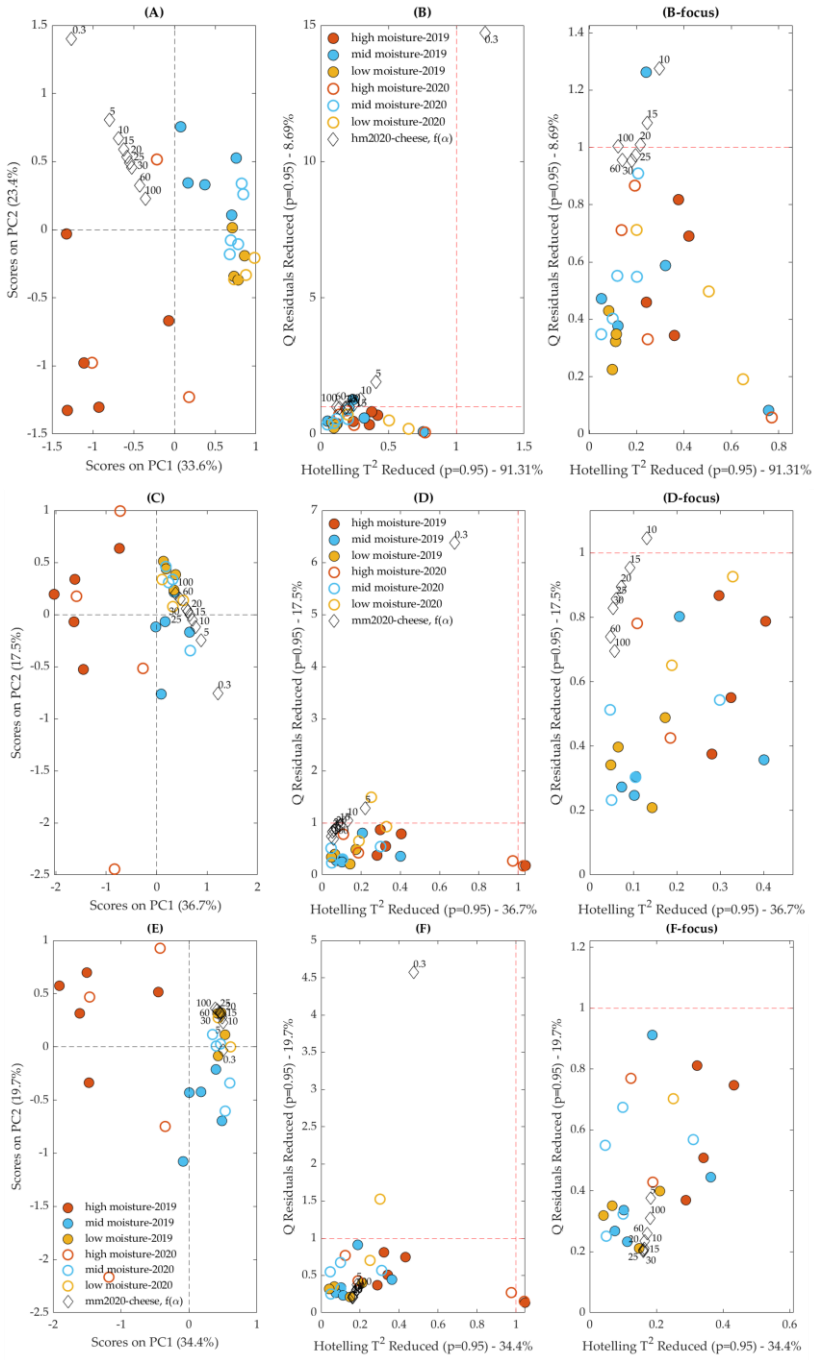


Figure 9.5. PCA Analysis. Projection of the Pecorino samples into the space of the first two PCs, highlighting the variation in the (PC1,PC2) coordinates for an high (A), mid (C) and low (E) moisture content cheese as side-effects of changing in α. Corrispective *Hotelling* T^2 *vs Q residuals* are displayed in plots (B), (D), and (F). 8, 6, and 6 PCs were retained for *hm*-, *mm*-, and *lm*-models, respectively.

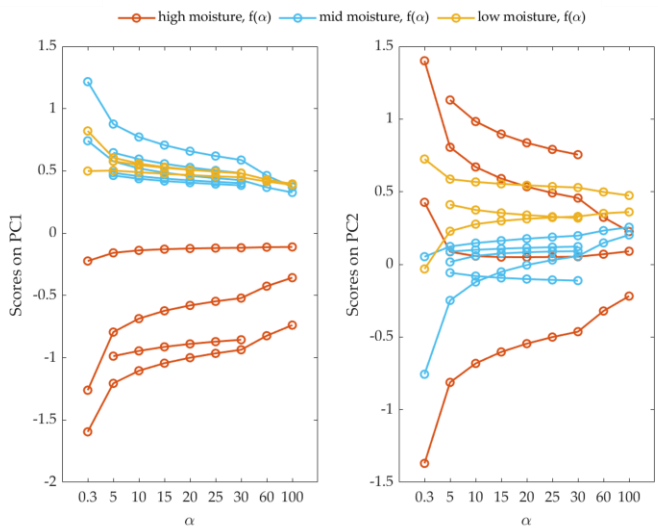


Figure 9.6. Evolution of the scores on PC1 and PC2 as response for the variation of the regularization parameter in the acquisition of relaxation data.

Looking at the Figure 9.5 and 9.6 the influence of the value of α parameter during the acquisition on the behavior of the objects in exploratory data analysis results evident. In some cases (i.e., for some values of α), the 2D spectrum cannot be even considered inside the model. Generally, the scores on PC1 tend to increase with an increasing of α for the *hm*-cheeses and to decrease for *mm*- and *lm*-samples; the scores on PC2 are also influenced but a systematic trend cannot be highlighted. Anyway, the standing point is the assessing of a relation between the α-value and the (PC1,PC2)-coordinates; further conclusions considering the sign of this relation are not drawn because of the rotation ambiguity of the principal component analysis.

We eventually decided to perform all the analysis at a fixed value of the smoothing parameter to be able to perform an effective (i.e., differences are due only to chemical-physical effects) comparison among the distribution functions of cheeses with different ripening time and acquired at different times. The coefficient α=20, standing at the center of the *plateau* in the relation PCs scores-α value, was considered to ensure more stable outcomes of the 2D-NMR-R measurements.

9.3.3 ASCA Analysis

The ANOVA-like partitioning was applied on the T_1/T_2 ratio projections from the 2D spectra as in equation 9.2; thirty-six (12 cheeses measured at three different times) samples were eventually available for this analysis. The data matrix was mean-centered prior to the application of ASCA. The SC Analysis was performed on the effect-matrices. 10000 data permutation rounds were performed to assess whether the effect of the two considered factors and their interaction can be considered statistically significant or not; the results are graphically displayed in Figure 9.7.

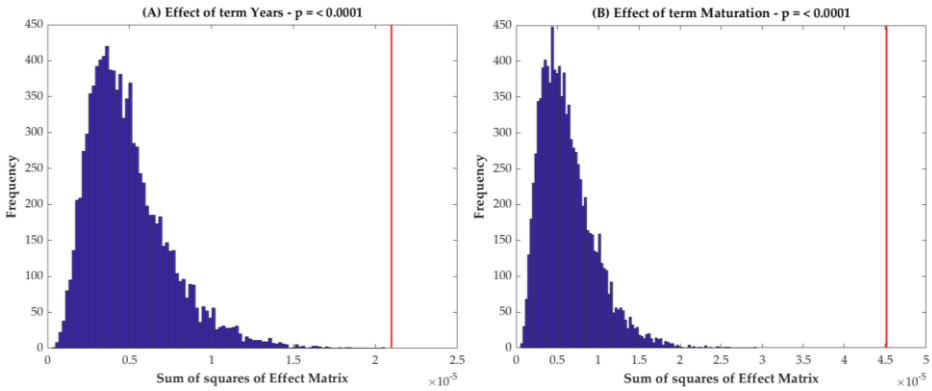


Figure 9.7. ASCA Analysis. Validation of the main effect of the factors Years (A) and Maturation (B): comparison between the value of the sum of squares obtained from the original design (red vertical line) and the distribution of the sum of squares of the null hypothesis (blue histogram) obtained by permutation test.

The explained variance for $\boldsymbol{X}_{Years}$, $\boldsymbol{X}_{Maturation}$, and $\boldsymbol{X}_{Years*Maturation}$ was 15.41%, 33.16% and 2.95% of the total data variance, respectively. For both factors *Years* and *Maturation* permutation test returned a p-value<<0.05, indicating time of storage and initial water content being significant effects. The two effect matrices $\boldsymbol{X}_{Years}$ and $\boldsymbol{X}_{Maturation}$ were afterwards decomposed as in eq. 3.22 and SCA scores accounting for within-level sample variability were estimated in accordance with Ref. 228:

$$\boldsymbol{T}'_{(\boldsymbol{X}_{effect}+\boldsymbol{E})} = (\boldsymbol{X}_{effect} + \boldsymbol{X}_{res})\boldsymbol{P}_{\boldsymbol{X}_{effect}} \qquad \text{(eq. 9.3)}$$

The loadings $\boldsymbol{P}_{\boldsymbol{X}_{Years}}$ and $\boldsymbol{P}_{\boldsymbol{X}_{Maturation}}$ were also investigated to evaluate which spectral features are mainly responsible for the differentiation of the samples stored for a different amount of time and with different water content.

In Figure 9.8 the plots of the back-projected scores and loadings for $\boldsymbol{T}'_{(\boldsymbol{X}_{Years}+\boldsymbol{E})}$ (Figure 9.8, A-B) and $\boldsymbol{T}'_{(\boldsymbol{X}_{Maturation}+\boldsymbol{E})}$ (Figure 9.8, C-D) are shown. The 95% confidence intervals

of the loadings have been estimated by a bootstrapping procedure consisting of 1000 repetitions. 79.41% of the variation was explained by the first component and 20.59% by the second one for the effect *Years*, whereas for the *Maturation* effect 89.76% and 10.24% are the portions of variability SC1 and SC2 accounted for, respectively.

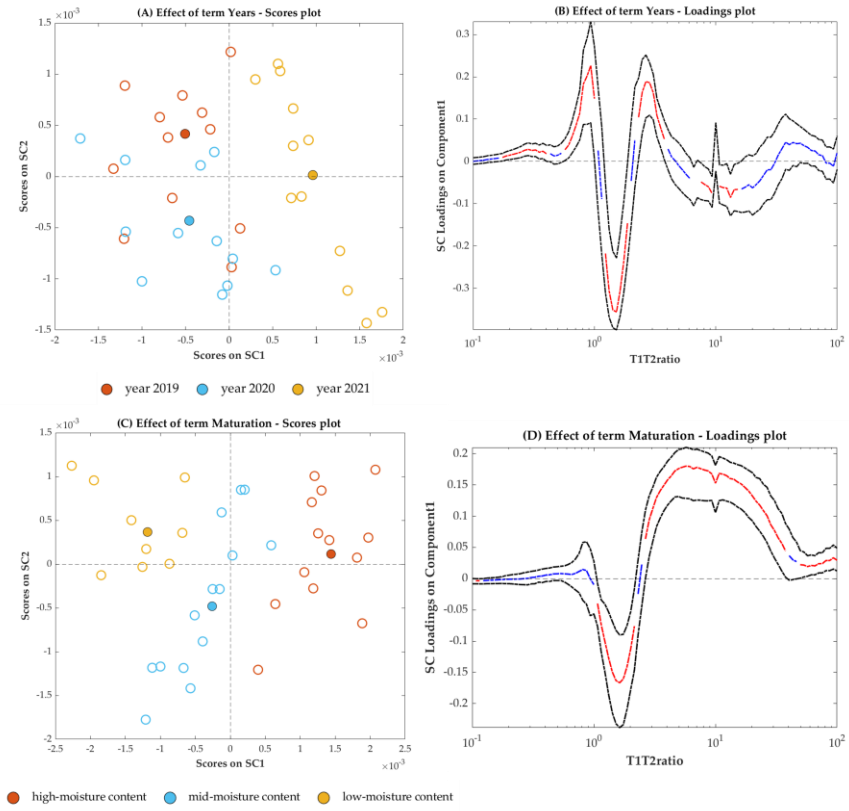


Figure 9.8. SCA modelling of the effects Years (A-B) and Maturation (C-D). Scores (A-C) plots. Full symbols represent the level averages (i.e., the centroid), empty symbols represent the residuals. Loadings plot (B-D). Loadings of the spectral variables on SC1 together with their bootstrapped confidence intervals (dashed black lines). Loadings are highlighted in red when significantly different from the zero, otherwise are highlighted in blue.

Looking at the Figure 9.8(A), a clear distinction of the 2021-acquisitions from the 2019 and 2020 groups, partially overlapping, can be visualized along SC1, supporting the result of the permutation test in the assessment of the effect significance. The relative distribution of the objects in the SCs-space of Figure 9.8(C) confirms also the significance of the effect *Maturation*, even in higher extent respect of the effect *Time*, since the distance among the factor level averages (fully symbols) is higher than the randomic dispersion among the samples due to the variability not accounted by the effect matrices. Also in

this case, SC1 is enough effective in displaying the dissimilarity among the three groups. To relate the observed trends with variations in the relaxation profiles, the loadings on SC1 can be inspected in Figure 9.8(B and D). Being the SC1 the most representative in both cases, loadings on SC2 are not displayed. The domain of liquid fat (higher absolute values of loadings in Figure 9.8(B)) is almost the only one affected in a significant extent by the aging of Pecorino cheeses over the three years of storage. Conversely, the whole network of fat and water in cheese is highly influenced by the different extent of maturation. Therefore, looking at the ASCA results, it seems that the proposed method is capable of recognize cheeses stored from different amounts of time, regardless of their different moisture content (key factor in food preservation), and *vice versa*.

The same analysis had been performed also on the unfolded final $\boldsymbol{X}$ (36x22500) matrix. The same results were obtained, but also the *Time*Maturation* interaction resulted to be significant in this latter case. Nevertheless, the drawback of this approach was the non-interpretability of the resulting loadings.

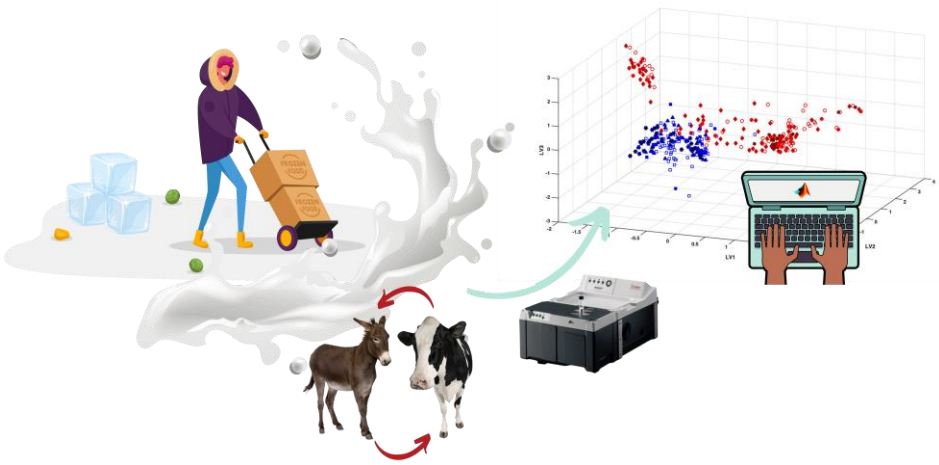
9.4 Conclusions and perspectives

In this work, 2D NMR-R Relaxometry coupled with EDA diagnostic tools and ASCA approach was employed as non-destructive, rapid, and green approach to characterize dairy commodities, using directly the information hid in the proton T_1-T_2 distribution functions. Fourteen Pecorino cheeses at different maturation levels were analyzed over three years, i.e., during their storage. An important goal of this research was first of all the application of a novel approach involving Principal Component Analysis in its diagnostic version for the optimization of a key parameter in the acquisition of NMR spin-echo decays. The influence of smoothing factor value on the behavior of Pecorino samples in the calculated benchmark models was indeed investigated. In the second part of the work, the effect of the storage-time and the moisture content (key factor in food preservation) were investigated by ASCA, focusing on the systematic variations into the NMR data set. This further analysis demonstrated the significant effect of both factors and the consequent ability to recognize cheese with different moisture content and time of storage with a portable technique, easy-to-implement *inline* in the food supply-chain.

The presented findings represent a solid starting point for future investigations. A further integration of the signal-density distributions over the full relaxation time range may result into a deeper understanding of the volume fraction occupied by the water. Moreover, an implementation of the proposed α-selection approach, involving cross-validation and increasing the number of samples, may produce data with higher robustness. Additionally, the ASCA approach might be modified and adapted to directly handle the 3D-array embedded in each $f(T_1, T_2)$ distribution.

Chapter 10

Near Infrared Spectroscopy coupled to chemometrics for the authentication of donkey milk



This project was carried out in collaboration with Prof. Federico Marini's research group (La Sapienza, Rome).

10.1 Introduction

Donkey milk is attracting growing interest in human nutrition as a valuable alternative in feeding infants because of its similarity with the human one and its hypoallergenic properties[270]. Several authors[271] have proved that pH, content in protein, ash and total solids in donkey milk is more similar to mare and human milk than those of all the other mammals. In addition, it stimulates the immune system, regulates the gastrointestinal flora, and prevents inflammatory diseases[272]. Coppola and collaborators suggested also

---

[270] Souroullas, K.; Aspri, M.; Papademas, P. *Food Res. Int.* **2018**, *109*, 416-425.
[271] a) Guo, H.Y.; Pang, K.; Zhang, X.Y.; Zhao, L.; Chen, S.W.; Dong, M.L.; Ren, F.Z. *J. Dairy Sci*. **2007**, *90*, 1635-1643; b) Salimei, E.; Fantuz, F.; Coppola, R.; Chiofalo, B.; Polidori, P.; Varisco, G. *Anim. Res*. **2004**, *53*, 67-78; c) Vincenzetti, S.; Polidori, P.; Mariani, P.; Cammertoni, N.; Fantuz, F.; Vita, A. *Food Chem*. **2008**, *106*, 640-649.
[272] Derdak, R.; Sakoui, S.; Pop, O.L.; Muresan, C.I.; Vodnar, D.C.; Addoum, B.; Vulturar, R.; Chis, A.; Suharoschi, R.; Soukri, A.; Khalfi, B.E. *Foods*. **2020**, *9*, 1302.

the use of donkey milk for probiotic purposes[273], providing a good growth medium for probiotic lactobacilli strains thanks to its high content of lysozyme and lactose. Despite its unique nutrient profile and economic potential, the sale of donkey milk has been allowed in Europe only in 2004, according to the EC Regulation n. 853/2004[274], which included it in the "other milk species" category. Nowadays, donkey breeding is circumscribed to Asia, Africa, eastern and some Countries of western Europe, and it is so dispersed that the milk yield is very low. Thanks to its properties and due to its niche-product status, donkey milk is a high-added value product, with a relatively high market value (around 15 € per liter), and, as such, it can be subjected to adulteration with milk from other mammalian species, especially with cheaper cow milk whose average European price at the barn has been 35.69€/100kg during the 2021[275]. This is turning into a common illicit practice, involving also other valuable kinds of milk, such as buffalo, goat and sheep milk; as a consequence, several approaches have been proposed in literature in order to detect and prevent these kind of frauds[276]. Nevertheless, there is still a lack of methods finalized to the authentication of donkey milk. These fraudulent practices result in an authenticity problem, but also in a real health issue. Indeed, recent clinical studies confirmed the feeding with donkey milk as the safest and the most valid alternative in nutrition of infants affected by Cow Milk Protein Intolerance (CMPI), an Ig-E mediated severe pathology affecting 3% of infants under the age of 12 months, often misdiagnosed[277].

The adulteration-detection study proposed here lays on these considerations and aims at developing a non-destructive approach for authenticating donkey milk and detecting its adulteration with cow milk. Near Infrared (NIR) Spectroscopy was employed to achieve this goal, because it had proved to perform effectively in such authentication issues[278]. NIR was used to analyze 67 samples of genuine donkey milk and 80 mixtures of donkey and cow milk (at different ratios) to mimic the adulteration. Moreover, both pure and adulterated samples were subjected to successive freezing-thawing (F/T) cycles prior to being re-analyzed by NIR. This physical sample pre-processing was included in order to investigate whether this pre-treatment alters milk specimens. Freezing products and

---

[273] Coppola, R.; Salimei, E.; Succi, M.; Sorrentino, E.; Nanni, M.; Ranieri, P.; Blanes, R.B.; Grazia, L. *Ann. Microbiol*. **2002**, *52*, 55-60.

[274] European Commission. Regulation (EC) No 853/2004. **2004**. *Specific hygiene rules for food of animal origin.* Retrieved from https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32004R0853 (last access January 2022).

[275] European Commission. Milk market observatory. Retrieved from https://ec.europa.eu/info/food-farming-fisheries/farming/facts-and-figures/markets/overviews/market-observatories/milk_en (last access January 2022).

[276] a) Agrimonti, C.; Pirondini, A.; Marmiroli, M.; Marmiroli, N. *Food Chem*. **2015**, *187*, 58-64; b) Pesic, M.; Barac, M.; Vrvic, M.; Ristic, N.; Macej, O.; Stanojevic, S. *Food Chem*. **2011**, *125*, 1443-1449.

[277] Ewing, W.M.; Allen, P.J. *Pediatr. Nurs*. **2005**, *31*, 486-493.

[278] a) Dos Santos Pereira, E.V.; de Sousa Fernandes, D.D.; de Araújo, M.C.U.; Diniz, P.H.G.D.; Maciel, M.I.S. *Microchem. J*. **2021**, *163*, 105885; b) Teixeira, J.L.D.; Caramês, E.T.D.; Baptista, D.P.; Gigante, M.L.; Pallone, J.A.L. *J. Food Compos. Anal*. **2021**, *96*, 103712.

selling them as fresh food is a further illegal action commonly applied[179,180], which has to be prevented[279].

Partial Least Squares-Discriminant Analysis (PLS-DA) was used to distinguish pure and adulterated donkey milk samples, before and after freezing-thawing procedure. The choice of the classifier fell on PLS-DA because it had proved to be a suitable tool for the authentication of dairies[66, 280] or in similar situations in combination with NIR spectroscopy[281].

Eventually, in order to check the significance of freezing-thawing treatment, data was handled by Analysis of Variance (ANOVA)-Simultaneous Component Analysis (ASCA).

10.2 Materials and methods

10.2.1 Donkey milk samples

Donkey milk samples were supplied directly from breeders settled in Abruzzo (Central Italy). Eighty samples were artificially adulterated mixing pure donkey and cow milk (purchased in local supermarkets). The sophistications were performed gradually increasing percentages of adulterant (from 2% up to 17%). Data was also used to attempt the quantification of cow milk content in donkey milk. One hundred and forty-seven samples were available in total, sixty-seven pure and eighty mixtures. Pure and adulterated samples were analyzed before (t0) and after undergoing one (t1), two (t2) and three (t3) freezing-thawing (F/T) cycles. In particular, forty pure and 80 adulterated samples were analyzed at t1, and 40 pure and 20 adulterated mixtures at both t2 and t3.

---

[279] Hassoun, A.; Shumilina, E.; Di Donato, F.; Foschi, M.; Simal-Gandara, J.; Biancolillo, A. *Molecules*. **2020**, *25(19)*, 4472.

[280] a) Aït-Kaddour, A.; Hassoun, A.; Bord, C.; Schmidt-Filgueras, R.; Biancolillo, A.; Di Donato, F.; Temiz, H.T.; Cozzolino, D. *Food Bioprocess Technol*. **2021**, *14*, 781-803; b) Genis, D.O.; Sezer, B.; Durna, S.; Boyaci, I.H. *Food Chem*. **2021**, *336*, 127699; c) Mabood, F.; Jabeen, F.; Ahmed, M.; Hussain, J.; Al Mashaykhi, S.A.A.; Al Rubaiey, Z.M.A.; Farooq, S.; Boqué, R.; Ali, L.; Hussain, Z.; Idrees, M.; Manzoor, S. *Food Chem*. **2017**, *221*, 746-750.

[281] a) Firmani, P.; La Piscopia, G.; Bucci, R.; Marini, F.; Biancolillo, A. *Microchem. J*. **2020**, *152*, 104339; b) Haughey, S.A.; Galvin-King, P.; Ho, Y.-C.; Bell, S.E.J.; Elliott, C.T. *Food Control*. **2015**, *48*, 75-83; c) Ríos-Reina, R.; García-González, D.L.; Callejón, R.M.; Amigo, J.M. *Food Control*. **2018**, *89*, 108-116.

10.2.2 NIR measurements

A Nicolet 6700 (Thermo Scientific Inc., Madison, WI) FT-NIR instrument equipped with an integrating sphere allowed the direct analysis of all the available samples without any further pre-treatment. A suitable volume of each sample was introduced into a glass vial placed on the window of the sphere and two analytical replicates were measured for each of them. The measurements were performed in reflectance mode in the spectral range between 4000 cm^{-1} and 10000 cm^{-1}, at a nominal resolution of 4 cm^{-1}. The signals were exported from the OMNIC software (Thermo Scientific Inc., Madison, WI) to the Matlab environment (R2020b; The Mathworks, Natick, MA) and converted to pseudo-absorbance (log(1/R)) to be further processed.

10.2.3 Statistical Analysis

In the present study, the PLS approach was used to quantify the cow milk content in the adulterated donkey milk based on NIR data, using Variable Sorting for Normalization (VSN) as data pre-processing technique. The reader is addressed to Appendix A[1] for more details on the algorithm of VSN. Moreover, the two-class problem pure/adulterated was coped with PLS-DA, so, the dummy $\boldsymbol{y}$ was a vector where each i-th object belonging to Class *pure* was represented by the value $y_i = 1$ whereas, an i-th sample belonging to Class *adulterated* was encoded as $y_i = 0$.

ASCA was instead applied in order to investigate the effect of the time factor (as F/T cycles) and its significance on the NIR-fingerprint of the pure donkey milk.

The mean-centered matrix $\boldsymbol{X}_c$ (of dimensions NxM), containing the NIR experimental data ($\boldsymbol{X}$), was partitioned into the matrices accounting for the effect of freezing/thawing ($\boldsymbol{X}_{F/T}$) and for the residuals ($\boldsymbol{X}_e$), associated to the random experimental error, according to the eq. 10.1:

$$\boldsymbol{X}_c = \boldsymbol{X}_{F/T} + \boldsymbol{X}_e \qquad \text{(eq. 10.1)}$$

By calculation, $\boldsymbol{X}_{F/T}$ will be a NxM matrix containing s (s: number of cells at t0 level) copies of the average vector calculated on all the samples for which factor F/T is at level t0, o (o: number of cells at t0 level) copies of the average vector calculated on all the samples for which factor F/T is at level t1, and so on. To assess the extent of the contribution of the factor, the Frobenius norm of the matrix corresponding to the effect is calculated and its comparison with the distribution of the values of the sum of squares related to the null hypothesis is performed, according to eq. 3.21. The distribution corresponding to the null hypothesis was estimated by means of a randomization of the group labels in the data. The spectral profile for a given sample remains unchanged, but the sample is randomly reassigned to one of the treatment groups (t0, t1, t2 and t3). Simultaneous Component Analysis is then computed on the effect matrix resulting from the ANOVA to model the variability associated to the effect of the freezing/thawing cycles.

The theory of the mentioned Chemometric techniques is described in Chapter 3. The Matlab functions employed in the present work can be freely downloaded from the RomeChemometrics website[221].

10.3 Results and Discussion

Prior to further investigations, the two replicates acquired on each sample were averaged, leading to the NIR profiles displayed in Figure 10.1.

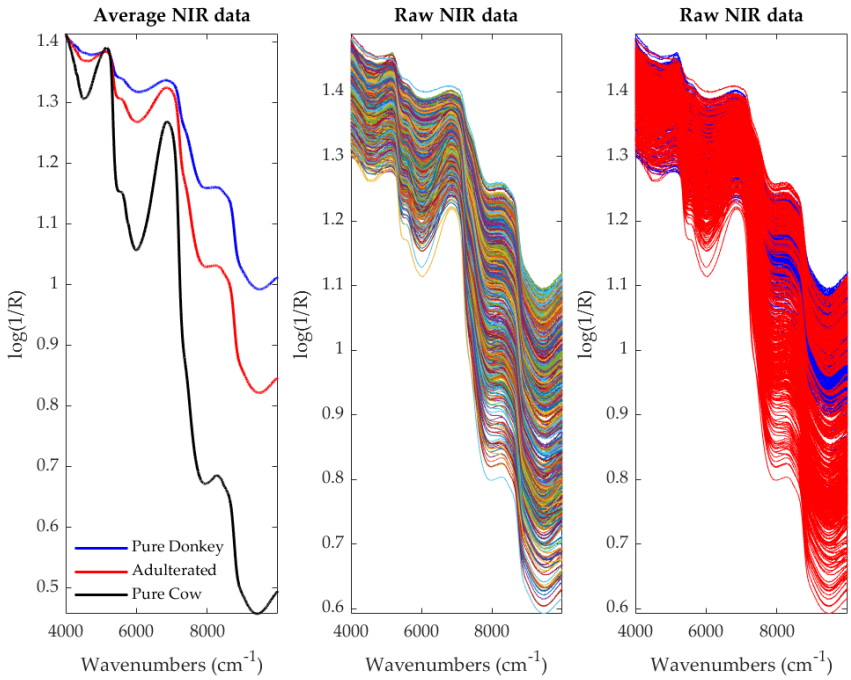


Figure 10.1. Near infrared spectra collected on the milk samples and comparison among the average signals collected for pure donkey milk, adulterated donkey milk and pure cow milk.

As above-mentioned, the aim of the present work is to develop a method exploiting the NIR spectroscopy to discriminate pure donkey milk and to quantify the amount of cow milk added as adulterant. In order to calibrate and validate the classification and the regression models, t0-data was split into two (training and test) representative sets of samples by the application of the Duplex algorithm on each class separately. The augmented matrix obtained by concatenating row-wise the NIR spectra differently pre-processed (raw data, SNV, 1st der., 2nd der., SNV+1st der., SNV+2nd der.) was subjected to

ComDim[253] analysis; three common components (CCs) were calculated and the two score-matrices ($\boldsymbol{T}_{pure}$ and $\boldsymbol{T}_{adulterated}$) obtained. Eventually, Duplex algorithm was run individually on these latter ones and the spectra were divided accordingly. Of the 147 t0-milk' samples, 112 (25 pure and 60 adulterated) were selected as training set and 35 (15 pure and 20 adulterated) were employed as test set for the validation of the models. All the t1, t2 and t3 samples were added to the test set in the attempt of applying the model calibrated on the fresh sample in the prediction of purity of the freezed-thawed ones.

10.3.1 Authentication of donkey milk: PLS-DA classification of milk samples

The most suitable data pre-processing approach was defined together with the optimal number of latent variables (LVs) in a 7-fold cross-validation procedure on the training samples looking at the accuracy as objective function to minimize; the results are shown in Table 10.1.

Table 10.1. Results in cross-validation of Partial Least Squares-Discriminant Analysis (PLS-DA) for the t0-NIR data set.

Pre-treatment	LVs	Accuracy (%)	Mean CCR (%)	Adulterated CCR(%)	Pure CCR(%)
Mean-centering (MC)	9	100.00	100.00	100.00	100.00
SNV+MC	**6**	**100.00**	**100.00**	**100.00**	**100.00**
1st derivative+MC	5	99.11	99.17	98.33	100.00
2nd derivative+MC	6	92.86	93.21	88.33	98.08
SNV+1st derivative+MC	5	99.11	99.17	98.33	100.00
SNV+2nd derivative+MC	4	91.96	92.24	88.33	96.15

In particular, the Standard Normal Variate[257] (SNV) was tested, alone or in combination with derivative methods, and the first and second derivatives were calculated using the Savitzky-Golay[258] filter with a 19 or 33 points windows and a 2nd or 3rd order polynomial fitting, respectively. Mean-centering was performed prior to the creation of each model as the normalization step. Looking at Table 10.1, it seems that the simple application of SNV (plus MC) is the most appropriate data pre-treatment because it guarantees the highest accuracy in cross-validation (100%) with the minimum number of LVs (six). Therefore, the PLS-DA model with six latent variables, which explained 99.7% of variance on $\boldsymbol{X}$-block and 45.4% on $\boldsymbol{Y}$-block, was retained. Successively, the optimal PLS-DA model was applied on the test samples providing 100% of correct classification for both pure and adulterated donkey milk classes. In Figure 10.2 the results of PLS-DA analysis are graphically represented in terms of calculated and predicted $\boldsymbol{y}$ values as a function of the training/test sample index.

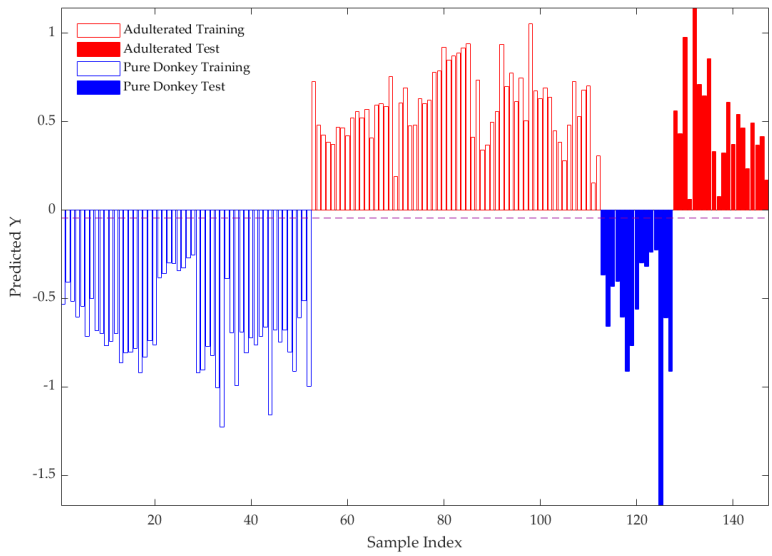


Figure 10.2. PLS-DA Analysis. Calculated (training set, empty-bars) and predicted (test set, full-bars) responses vs sample index. The red dashed line is the threshold dividing the two class-regions.

Variables Importance in Projection (VIP) indices were calculated for a deeper investigation of the relevancy of each spectral variable in the definition of the latent variables space. According to the *greater-than-one* criterion, variables with VIP value higher than 1 (highlighted in red or blue in Figure 10.3) are considered contributing the most to the model. As a further tool in the chemical interpretation of the model, a distinction between the spectral variables with VIP score>1 and regression coefficient>0 (red in Figure 10.3) and VIP score>1 and regression coefficient<0 (blue in Figure 10.3) can be done. The intensity of variables highlighted in red are found to be higher in the adulterated class, on the contrary the intensity of blue-variables in Figure 10.3 is higher in the pure class.

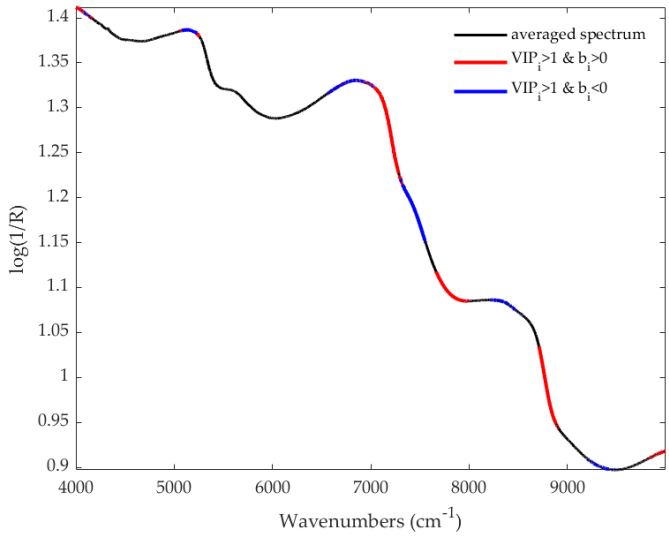


Figure 10.3. PLS-DA Analysis. Graphical representation of the spectral variables presenting VIP index greater than 1. The average NIR spectrum is represented in black, the variables with (VIP>1 and regression coefficient>0) are highlighted in red and the variables with VIP>1 and regression coefficient<0 in blue.

By inspection on Figure 10.3, the spectral ranges which appeared as the most discriminant are mainly associated to moisture and fats content[282]; in particular, these are:

- the spectral range between 4000 cm^{-1} and 4200 cm^{-1}, probably associable to the combination bands of C–H and C–O stretching vibrations in fats;
- the variables between 5050 and 5250 cm^{-1} ascribable to the moisture, probably dependent on the different water content in donkey and cow milks;
- a large area centered at 7020 cm^{-1};
- a large area centered at 8400 cm^{-1} where the absorptions are mainly due to the second overtone of the C–H stretching in fats;
- around 9200 and 10000 cm^{-1}.

Eventually, the optimal PLS-DA model (SNV+MC, 6 LVs) was used to predict the class-membership of the samples after the different F/T cycles. The results are displayed in

[282] Ayvaz, H.; Mortas, M.; Dogan, M.A.; Atan, M.; Yildiz Tiryaki, G.; Karagul Yuceer, Y. *J. Food Sci. Technol.* **2021**, *58(10)*, 3981–3992.

Table 10.2, in terms of total classification accuracy, mean correct classification rate and class-sensitivities and specificities, and graphically in Figure 10.4.

Table 10.2. Results in prediction (on t0-, t1-, t2- and t3-data) of the optimal PLS-DA model trained on the t0-NIR data set. All the figures of merit are expressed as percentages.

	LVs	Accuracy	Mean CCR	Adulterated Sensitivity	Pure Sensitivity
t0	6	100.00	100.00	100.00	100.00
t1	6	85.71	84.78	87.50	82.05
t2	6	79.66	84.62	100.00	69.23
t3	6	79.66	84.62	100.00	69.23

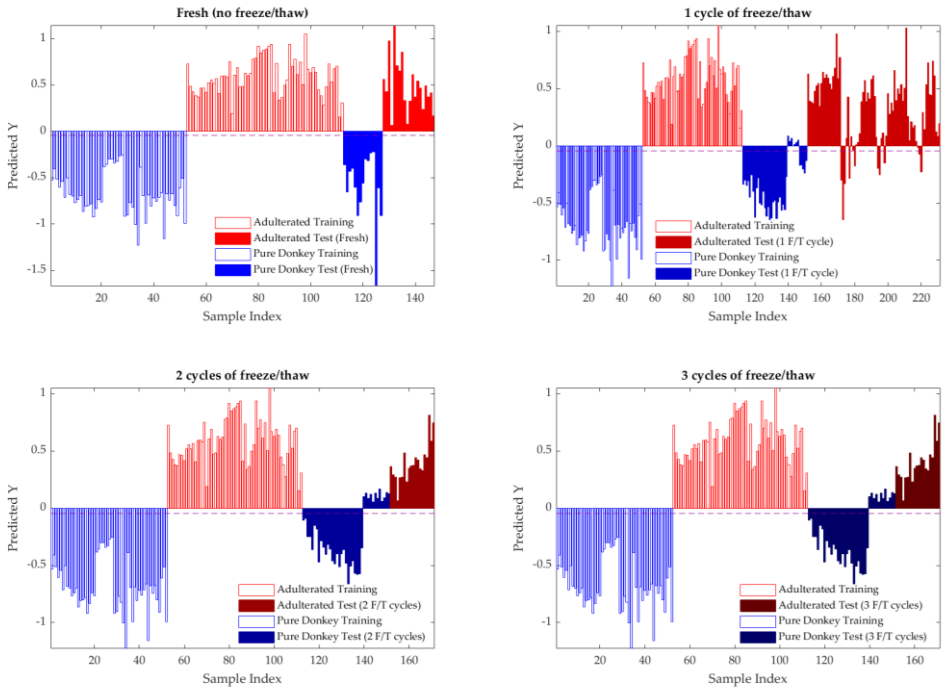


Figure 10.4. PLS-DA model application on t1 (1 cycle of freeze/thaw), t2 (2 cycles of freeze/thaw) and t3 samples (3 cycles of freeze/thaw). The red dashed lines are the thresholds dividing the two class-regions.

Among these further models, the highest accuracy was obtained including in the validation set only the samples which underwent a single F/T cycle (t1). When the spectra collected at the further time points (t2 and t3) are included in the test set, the model

achieves the same results, i.e., an accuracy of 79.66%. In general, it is possible to observe that the developed model efficiently classified pure and adulterated samples, slightly decreasing in accuracy with the increase of the freezing-thawing cycles, revealing a good robustness with respect to the F/T process.

10.3.2 Quantification of cow milk content in donkey milk samples

After having assessed the potentiality of the NIR spectroscopic fingerprint in the discrimination of adulterated donkey milk samples, the possibility of calibrating a model for the prediction of the concentration of cow milk as adulterant was investigated. To achieve this, a multivariate calibration model by the PLS algorithm was built using the t0-spectroscopic data of the training set (same for PLS-DA) pre-processed with VSN plus weighted-SNV. 7-fold venetian-blind cross-validation (c.v.) was used to choose the optimal model complexity (in order to avoid overfitting), using the root mean squared error of c.v. (RMSECV as evaluation parameter. The optimal model (5 latent variables) had a correlation coefficient of actual versus predicted in calibration in cross-validation of 0.75 and a RMSECV=0.039. Eventually, the model was applied on the test set achieving a RMSEP<0.040. Figure 10.5 shows the plot of predicted versus actual concentrations of cow milk in donkey milk, both for calibration (red) and external samples (black). The inspection of the figure reveals the non-complete suitability of the approach for an accurate quantification of the adulterant.

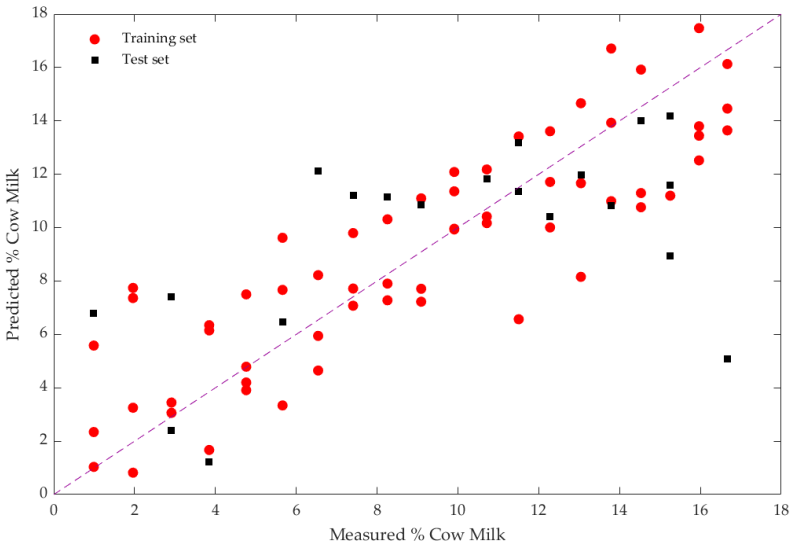


Figure 10.5. PLS regression. Predicted percentage of cow milk in the different mixtures with donkey milk. See Section 10.2.1 for more details of composition.

10.3.3 ASCA Analysis

The ANOVA-like decomposition was applied on the NIR spectra of the pure samples after SNV or VSN plus weighted-SNV correction and mean-centering. The design is balanced because the different levels (t0, t1, t2 and t3) of the considered factor (time as freezing/thawing cycles) have the same number of replicates (40 samples).

10.3.3.1 Estimation of the effect and its significance

In the condition, as in the present case, of mutual orthogonality of the effect estimates, the variation in $\|\boldsymbol{X}_c\|^2$ can be split in independent parts as described in the eq. 10.1:

$$\|\boldsymbol{X}_c\|^2 = \left\|\boldsymbol{X}_{F/T}\right\|^2 + \|\boldsymbol{X}_e\|^2 \qquad \text{(eq. 10.2)}$$

The different contributions in the case of the data set considered in this study are the term accounting for the freezing/thawing transformation ($\boldsymbol{X}_{F/T}$) and the residual variation ($\boldsymbol{X}_e$), i.e., the random variation among the replicates that is not accounted by the model. In particular, the amount of variance explained by the terms $\boldsymbol{X}_{time}$ and $\boldsymbol{X}_e$ were 61% and 39% for the SNV-model and 6% and 94% for the VSN ones, respectively, revealing that the large effect of the F/T procedure on the NIR fingerprint could be more efficiently highlighted using the SNV pre-treatment. To assess if the difference of the effect from the zero is statistically significant (not a sampling fluctuation), a procedure based on permutation test was adopted, as described in Section 10.2.3. The empirical distribution of the values of the sum of squares (SSQ) corresponding to the null hypothesis was estimated using 10000 randomizations. The results obtained in both cases, i.e., ANOVA-like decomposition on centered data with SNV and VSN plus weighted-SNV correction, are displayed in Figure 10.6.

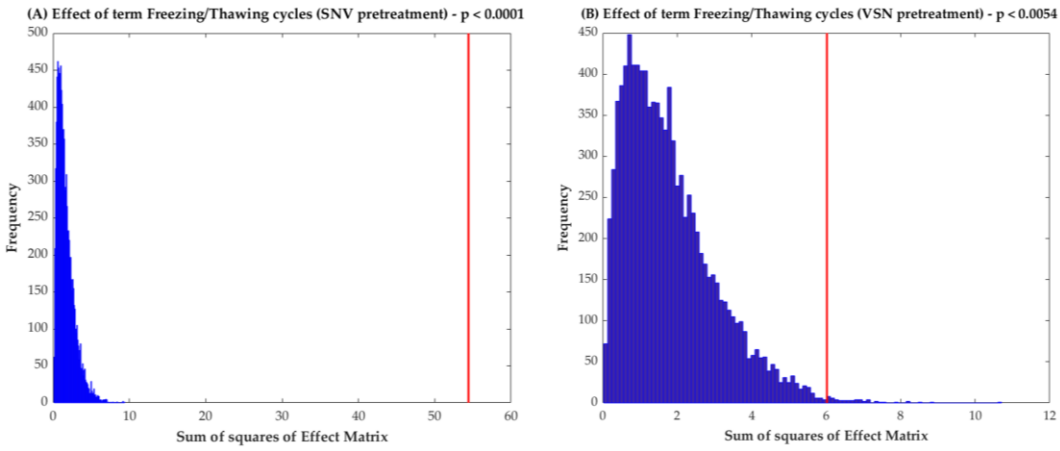


Figure 10.6. ASCA Analysis. Validation of the main effect of the factor Freezing/Thawing Cycles: comparison between the value of the sum of squares obtained from the original design (red vertical line) and the distribution of the sum of squares of the null hypothesis (blue histogram) obtained by permutation test; (A) SNV pre-treatment, (B) VSN pre-treatment.

The SSQ obtained from the partitioning of the data matrix according to the original design is larger than the threshold value. Therefore, it can be concluded that the effect of freezing/thawing cycles results to be statistically significant (p(SNV)<0.0001, p(VSN)<0.0066) for whatever pre-treatment is applied, i.e., the F/T transformation affects the NIR spectral fingerprint of pure donkey milk.

10.3.3.2 SCA Models

SC Analysis was applied to extract the systematic information from the effect matrix derived from the ANOVA-decomposition. The factor *time* was studied at four levels (t0, t1, t2 and t3), therefore, three components were fitted. Also in this case, the modified score plot derived by Zwanenburg and co-workers was built, i.e., the residual term is back-projected ($\boldsymbol{T}'_{F/T}$) onto the principal component subspace spanned by the effect matrix ($\boldsymbol{P}^T_{F/T}$).

Figure 10.7 shows this new score plot for the SNV and VSN pre-treated NIR spectra. 98.06% of the variation was explained by the first component, 1.38% by the second one and the third one described only the remaining 0.56% for the SNV-data. For the VSN-data 80.61%, 18.47% and 0.92% of total variation were explained by the first, second and third component, respectively.

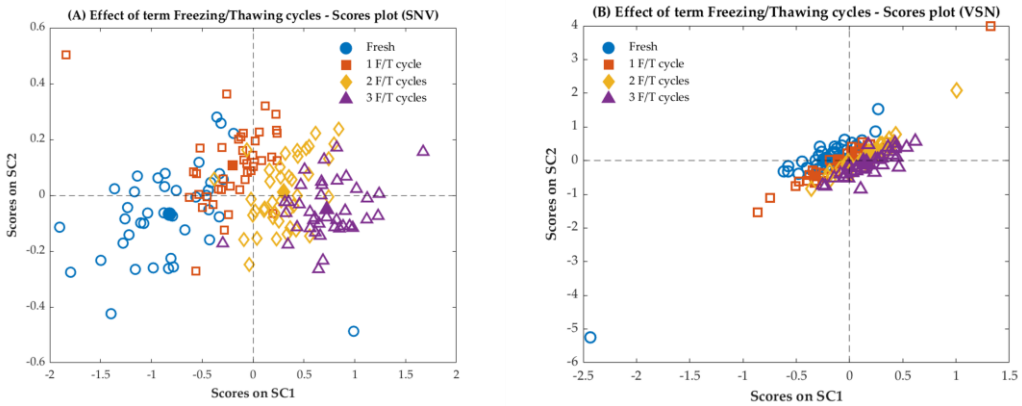


Figure 10.7. SCA modelling of the effect of F/T cycles ((A) SNV pre-treatment, (B) VSN pre-treatment). Scores plots. Full symbols represent the level averages, empty symbols represent the residuals.

The relative distribution of the objects in the space confirms the significance of the effect since the distance among the points corresponding to the factor level averages is higher than the dispersion among the replicates due to the natural variability. To relate the observed trends with variations in the spectral signal, the loadings on SC1 are reported in Figure 10.8.

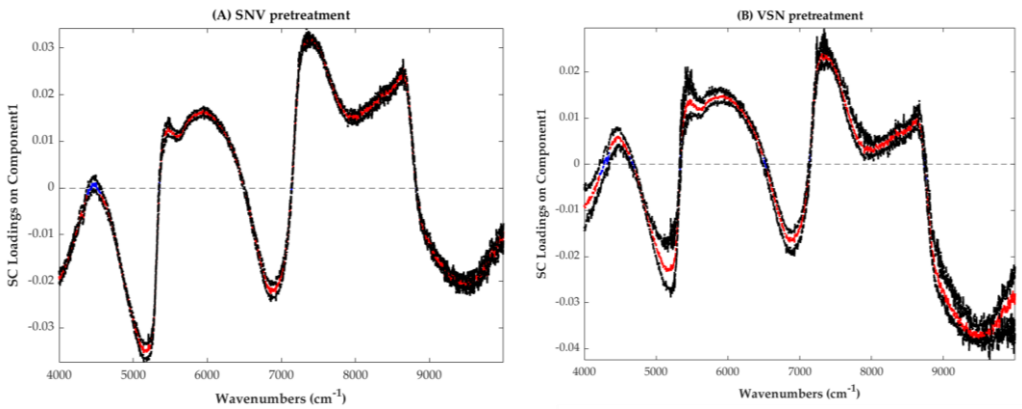


Figure 10.8. SCA modelling of the effect of F/T cycles ((A) SNV pre-treatment, (B) VSN pre-treatment). Loadings of the spectral variables on SC1 together with their bootstrapped confidence intervals (dashed black lines). Loadings are highlighted in red when significantly different from the zero, otherwise are highlighted in blue.

A bootstrap procedure was employed to estimate their confidence intervals at 95%. Visual inspection of the figure shows that the whole NIR-spectrum is affected by the variation in the "freshness" of the milk, additional F/T cycles will result in an increase of the intensity of signals with positive loadings and in a decrease of spectral signals associated with negative ones.

10.4 Conclusions

In this work, NIR Spectroscopy coupled with PLS-DA classifier was employed as a non-destructive, rapid, and green approach to authenticate a high added-value product as donkey milk, and to detect its possible adulteration with cow milk. The focus of the classification was not only on the discrimination between pure and adulterated specimens, but also on investigating whether the developed approach is able to pursue its purpose when samples have undergone different F/T cycles. As expected, the highest performances (100% accuracy on the validation set) were obtained classifying t0 samples. After the inclusion (in the test set) of the spectra collected at the other time points (t1, t2 and t3), the PLS-DA model demonstrated to be robust, only slightly decreasing its predictive capability from 85.71% to 79.66%.

In the second part of the work, the effect of a usual technological practice as freezing/thawing was investigated by ASCA. This approach was used to analyze the systematic variations into the spectral data set of pure samples. This further analysis demonstrated the F/T treatment has a statistically significant effect and that it spreads its impact on the whole NIR-fingerprint, as suggested by the Simultaneous Component Analysis.

These results indicate that the proposed method is suitable to be applied for the quality control of donkey milk samples and potentially for the quantification of cow milk as adulterant, irrespective of the conditions used for its storage.

Final Remarks

During the development of this doctoral thesis, I had the chance to grow personally and professionally inspired by mentors so dear to me. They gave me the possibility to prove myself bringing me into new paths, even distant from what I would be acquainted to. Many accomplishments I have reached thanks to this long journey. The last one is this thesis, in which I have put all my efforts.

In the research which constitutes the present PhD thesis, new analytical methods to handle different food traceability and authentication issues, focusing on dairy sector, have been addressed. More in details, the reader has been led through two Sections dealing with different food-related aspects.

Section 1 is a straightforward analysis of the implications and consequences coming from the selection of an experimental set-up/approach among all the possible ones. More in details, a geographical origin-cheesemaking process classification problem concerning the characterization and discrimination of three typical Pecorino cheeses produced in Central Italy and certified with European (PDO) and International (Slow Food Presidia) quality schemes is addressed. **Chapters 4-6** describe the use of single-instrument approaches belonging to the family of the *targeted/convoluted* techniques, whereas the efficiency of a *fingerprinting/untargeted* characterization is tested in **Chapters 7**. Overall conclusions based on the application of data fusion approaches are drawn in **Chapter 8**, giving a new point of view about the main topics of currently food research. Untargeted techniques are nowadays an invaluable resource in food traceability and authentication making possible fast, low-cost, and easy-to-handle applications. Nevertheless, their potential power can be fully expressed only when a large data set of samples is available, allowing to develop reliable and robust predictive models. This is a real challenge for the current times: we have the instrumentations and the skills to perform economic- and environmental-compliant daily controls in the food supply-chain (avoiding the random controls using the *convoluted* technologies), but any effort in this direction will be meaningless without the institution, promoted by the competent authorities, of intensively campaigns for creation and updating of largely food products-data bases.

Section 2 describes two different green, fast, and non-destructive approaches based on *fingerprinting* spectroscopic techniques, proposed with the aim of characterizing *inline* niche dairy products.

In **Chapter 9**, 2D NMR-R Relaxometry coupled with Chemometrics was employed as non-destructive, rapid, and green approach to characterize dairy commodities from Central Italy, using directly information hidden in the proton T_1-T_2 distribution functions. Fourteen Pecorino cheeses at different maturation levels were analyzed during their storage over three years. The effect of the storage/time and the moisture content (key factor in food preservation) were investigated, focusing on the systematic variations into the NMR data. This analysis demonstrated the ability of recognizing cheeses with different moisture content and time of storage by using a portable technique, easy-to-implement *inline* in the food supply-chain. Moreover, a novel approach for the optimization of a key parameter in the acquisition of NMR spin-echo decays is proposed. This project was carried out in collaboration with Prof. Nicoletta Spreti and Prof. Cinzia Casieri's research group (University of L'Aquila) and is currently not disclosed. The presented findings, indeed, represent a solid starting point for future investigations.

In **Chapter 10** the bulls-eye has been focused on donkey milk, an emerging food commodity which is gaining increasing attention due to its nutritional properties. The effect of the usual technological practice of freezing/thawing (F/T) was investigated by analyzing the systematic variations into the NIR-spectral data of milk samples. This analysis demonstrated the F/T treatment has a statistically significant effect and that it spreads its impact on the whole NIR-fingerprint. Moreover, a common dilution case is faced. As a high-added value product, indeed, donkey milk can be subjected to adulteration with cheaper milks. In this chapter the possibility of developing a fast and economic method for its authentication by means of Near Infrared (NIR) Spectroscopy coupled with Chemometrics is successfully proposed. This project was carried out in collaboration with Prof. Federico Marini's research group (La Sapienza, Rome).

Appendices

Nomenclature and abbreviations

Appendix A	
x_{ij}	generic element of the data matrix $\boldsymbol{X}$
$\boldsymbol{x}_i$	generic row *i* of the data matrix $\boldsymbol{X}$
$\bar{x}_j$	average value of the *j-th* variable
σ_j	standard deviation of the *j-th* variable
$\hat{x}_{ij}$	corrected element x_{ij}, column-pre-treatments
$\bar{\boldsymbol{x}}_i$	mean value of the original row *i* (e.g., spectrum)
σ_i	standard deviation of $\boldsymbol{x}_i$
$\hat{\boldsymbol{x}}_i$	corrected row (e.g., spectrum), row-pre-tratments
$\|\boldsymbol{X}\|_F$	Frobenius norm of matrix $\boldsymbol{X}$
$\boldsymbol{X}^T\boldsymbol{X}$	cross-product, variance-covariance matrix
$\boldsymbol{W}$	diagonal weighting matrix (MxM) in VSN with $w_{ij} = 0$ and $0 \leq w_{ii} \leq 1$
$U(x)$	generic polynomial expression in Savitzky-Golay (S.-G.) approach
u_{i+s}	enter of the set of uniformly spaced digital data points in S.-G.
$(p_s^{(q)})$	set of coefficients determined in S.-G.
q^{th}	smoothed value of the derivative in S.-G.
Appendix B	
$RMSEC$	Root Mean Square Error in Calibration
$RMSECV$	Root Mean Square Error in Cross-Validation
$RMSEP$	Root Mean Square Error in Prediction
n	number of observations in the considered set
y	real value of the response
$\hat{y}$	response calculated/predicted by the model
TP/TN	number of P and N samples correctly classified as P and N
FP/FN	number of P and N samples wrongly classified as P and N
c_{gg}	diagonal element of the confusion matrix
ng	number of individuals in the class g of the considered set
$CCR, \%$	Correct classification rate
$Acc, \%$	Accuracy

Appendix A

A1 Data pre-treatment

Very often, in data, relevant chemical information could be covered by unwanted effects (e.g., non-linear instrument responses, shift problems, misaligned baselines, light scattering, or other artefacts) of larger magnitude[283]. To reach an accurate model (both un- and supervised), data pre-processing has to be applied.

Following the classification proposed by Trygg and collaborators[284], data pre-processing methods can be generally divided into three *loosely* defined categories. The first category collects methods for denoising and filtering, generally referred to as signal-to-noise ratio (SNR) enhancing techniques. They are usually applied (in an *unsupervised* manner) before modeling of data to reduce the influence of the random variation without altering the signal. The classical *unsupervised* methods for spectral normalization and differentiation belong to the second category. Standard normal variate transformation and first and second derivatives are examples. The third category consists of model-based pre-processing methods. These methods for variable selection and dimensionality reduction are aimed at removing the not-relevant for $\boldsymbol{Y}$ systematic variation present in $\boldsymbol{X}$. They involve a training set for estimating parameters

Among these categories, many different pre-processing approaches have been proposed in literature; the brief description carried out in this appendix will be limited to those that have been chosen within the studies reported in the present thesis. For the sake of clarity, the two following paragraphs will deal with two slightly different aspects of data pre-treatment. First of all, a brief description of data pre-processing intended to make the signals recorded on many samples comparable by removing the non-informative variance related to offset differences (*centering*) or due to different units of measurement (*scaling*), strictly required before any data analysis, will be addressed. Secondarily, the pre-processing techniques not strictly necessary prior to data analysis and normally applied with the aim of limiting the undesired chemical and physical variations will be described.

---

[283] Rinnan, Å.; van den Berg, F.; Engelsen, S.B. *Trends Analyt. Chem.* **2009**, *28*, 1201-1222.

[284] Trygg, J.; Gabrielsson, J.; Lundstedt, T. In Brown, S.D.; Tauler, R. and Walczak, B. (Eds.) *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Volume 4 (pp. A1-A8), Amsterdam, The Netherlands: Elsevier, **2009**.

A1.1 Centering and scaling

Centering is normally applied to handle the offset in data, i.e., a part of the model that is constant across one or several modes. Offsets can indeed be constant across the first mode (rows), the second one (column), and both modes. If the data are centered by subtracting the column average from every element in the column, this is referred to as *centering across the first mode*[285], more often as *mean-centering*. More in details, it consists of subtracting from each element of the data matrix (x_{ij}) the average value of the corresponding *j-th* variable ($\bar{x}_j$), as shown in eq. A1:

$$\hat{x}_{ij} = x_{ij} - \bar{x}_j \qquad \text{(eq. A1)}$$

Offsets across the first mode are thus removed. From a geometrical point of view, removing the offset from each variable means to shift the center of the coordinates system (i.e., the zero) in the data centroid. The variation of each variable is thus moved to the zero-level. The effect of mean-centering is easily verifiable working in the PC space: if the data were not centered, the first principal component would describe only the distance between the zero and the centroid of the data, covering interesting information.

Scaling is an useful operation so that each variable has the same initial standard deviation (removing different measurement units). The traditional a*utoscaling* consists of both mean-centering and *scaling within the second mode* (i.e., each column is multiplied by a certain scalar):

$$\hat{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \qquad \text{(eq. A2)}$$

where σ_j is the standard deviation.

It has the effect, unlike mean-centering, to change the distances among samples in the variables (or PCs) space. Autoscaling amplifies the importance of low variance variables but has the drawback of amplifying even the noise; because of this, it is rarely used to pre-treat spectroscopic data.

The effect of mean-centering and autoscaling is effectively shown in Figure A1.

---

[285] Bro, R.; Smilde, A.K. *J. Chemom*. **2003**, *17*, 16-33.

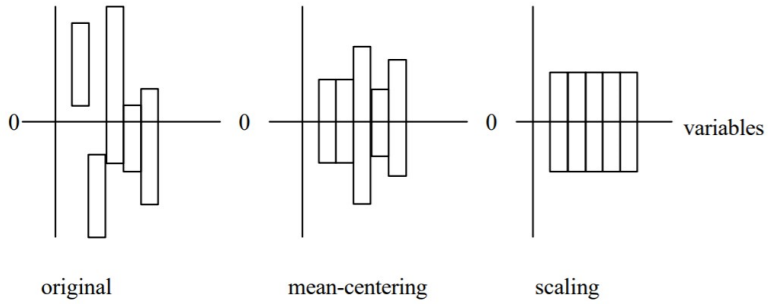


Figure A1. Effect of mean-centering and autoscaling.

When dealing with more-than-one matrix to be merged in a single dataset, as in the data fusion strategies, a *block scaling* operation needs often to be performed, if the algorithm requests it, to make the blocks comparable and to avoid the model being driven only by the blocks with the greater number of variables. Many block-scaling factors have been proposed[160]. One of the most employed is the use of the *Frobenius' norm* ($\|\boldsymbol{X}\|_F$), i.e., the total sum of squares of the matrix:

$$\widehat{\boldsymbol{X}} = \boldsymbol{X}./\|\boldsymbol{X}\|_F \quad \text{where } \|\boldsymbol{X}\|_F = \sqrt{trace(\boldsymbol{X}^T\boldsymbol{X})} \qquad \text{(eq. A3)}$$

A1.2 Others pre-processing techniques

Analytical signals, i.e., the responses of the instruments, are rarely submitted *as they are* to chemometric elaboration. Mathematical corrections, specific for addressing different situations, are indeed applied to minimize the contribution of undesired variations, either random or systematic, which are not useful for the characterization of the system under study. Such pre-treatments are usually applied independently on the vectors of signals arising from a single sample, so they are often referred to as *row pre-processing*.

They can have the purpose of minimizing *random* noise (averaging methods, polynomial smoothing, and so on) or correcting the *systematic* unwanted signal variations. Horizontal shifts (solved by alignment methods[286]), baseline vertical shifts (offsets) and drifts (slope variations), and global intensity effects (i.e., a global amplification or a global decreasing) are the most common of these systematic effects.

Baseline vertical shifts and global intensity variations are usually referred to as additive and multiplicative effects, unrelated to the quantity of interest, respectively. This means that, for a given value of the response y, $a\boldsymbol{x} + b$ is measured instead of $\boldsymbol{x}$, with $a \neq 1$ and $b \neq 0$. The multiplicative effect can be due to the quantity of product analyzed in

[286] Nielsen, N.-P.V.; Carstensen, J.M.; Smedsgaard, J. *J. Chromatogr. A.* **1998**, *805*, 17-35.

chromatography (*shape effect*[287] against *size effect*) or to the geometry of the measuring device (signal collected as reflectance) and the particle size of the product (signal expressed as absorbance) in spectroscopy[288]. The additive effect is a concept related to the *baseline*, that are assigned to the *background* in chromatography and NMR, or can be more closely intertwined with the useful portion of the signals in spectroscopy, because they come from the interaction between light and matter[289].

To deal with these issues, many signal transformations have been proposed. Among them, *Standard Normal Variate* (SNV) and the family of Multiplicative Scatter Correction (MSC) methods are the most used. Among these, SNV is one of those aligning the signals without the need for a reference spectrum.

As mentioned above, a spread issue in spectroscopy is the observation of differences between the measurements not only due to the chemical composition of the analytes, but also to light scattering. This phenomenon is related to the physical nature of the particles which constitute the sample (in solid form or in dispersion state) and to the non-constancy of the optical path in the sample (related to the particles size). For this reason, spectra show usually high collinearity, and additive (baseline vertical shifts) and multiplicative (global intensity variation)[290] effects unique for every spectrum.

---

[287] Filzmoser, P.; Walczak, B. *J Chromatogr A*. **2014**, *1362*, 194-205.
[288] a) Geladi, P.; MacDougall, D.; Martens, H. *Appl Spectrosc*. **1985**, *39(3)*, 491-500; b) Isaksson, T.; Næs, T. *Appl Spectrosc*. **1988**, *42(7)*, 1273-1284.
[289] Yang, L.; Miklavcic, S.J. *JOSA A*. **2005**, *22(9)*, 1866-1873.
[290] Helland, I.S.; Næs, T.; Isaksson, T. *Chemometri. Intell. Lab. Syst.* **1995**, *29(2)*, 233-241.

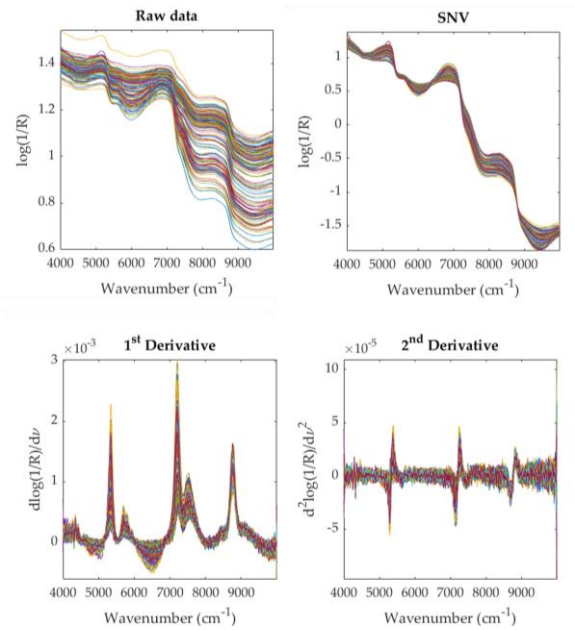


Figure A2. Raw and differently pre-processed (SNV, 1st der., and 2nd der) near infrared spectra. The data has been collected on the milk samples as part of the study described in Chapter 10.

To overcome this problem, SNV transformation has been proposed by Barnes[257] in 1989. SNV could be classified as a *row*(sample)-*wise* pre-treatment; it removes the slope variation on the i^{th}row of the data matrix as in eq. B4:

$$\hat{\boldsymbol{x}}_i = \frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}_i}{\sigma_i} \qquad \text{(eq. A4)}$$

where $\hat{\boldsymbol{x}}_i$ is the corrected spectrum, σ_i is the standard deviation of $\boldsymbol{x}_i$, and $\bar{\boldsymbol{x}}_i$ is the mean value of the original spectrum. The effect of SNV is shown in Figure A2.

SNV, like others pre-processing tools (e.g., derivatives), modifies the visual shape of signal profiles influencing, often radically, the significance of variables in the data matrix. This may have important implications when the role of such variables is investigated in the further unsupervised and supervised pattern recognition analysis: the *loading paradox*. In SNV-transformed spectra, spectral information is spread along the whole spectral range, with the risk of enhancing artificially not-actually significant variables[291]. This problem, not always verified, is very sneaky for non-selective fingerprinting signals.

[291] Oliveri, P.; Malegori, C.; Simonetti, R.; Casale, M. *Anal. Chim. Acta*. **2019**, *1058*, 9-17.

A novel algorithm named *Variable Sorting for Normalization* (VSN) has been proposed by Rabatel and collaborators[292]. This algorithm is built to produce automatically a weighting function, i.e., the diagonal matrix $\boldsymbol{W}$ (MxM) with $w_{ij} = 0$ and $0 \leq w_{ii} \leq 1$, to be introduced in SNV pre-processing, i.e., mean($\boldsymbol{xW}$) used as offset and std($\boldsymbol{xW}$) as denominator, favoring signal variables that are only impacted by additive and multiplicative effects, and not by the response(s) of interest. This allows significantly improving signal shape and model interpretation. More in details, the VSN method proposes to calculate w_{ii} as the probability that the variable i in the $\boldsymbol{X}$ matrix belongs to the group Sz, indicating a pure size effect. To evaluate this probability, subsequent random samplings of pairs of signals in the $\boldsymbol{X}$ matrix are performed and, for each of these pairs of profiles, variables belonging to Sz are identified. The probability w_{ii} is then estimated as the frequency of assignment of the variable i in the class Sz. Figure A3 represents a real example of the application of VSN.

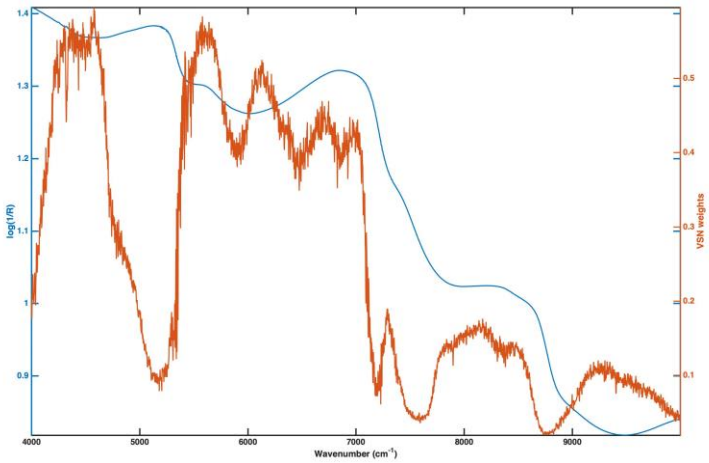


Figure A2. VSN weighting function. Near infrared raw profile (blue) and w_{ii} calculated for each wavenumber by VSN algorithm. The data has been collected on the milk samples as part of the study described in Chapter 10.

Focusing on baseline vertical shifts (offsets) and drifts (slope variations), numerical derivatives represent an efficient correction tool. In particular, first derivative corrects for baseline shifts, while second derivative corrects for both shifts and drifts. Moreover, derivatives may enhance minor structural differences among very similar signal profiles. In order to overcome the simultaneous decreasing of signal-to-noise ratio, derivation

[292] Rabatel, G.; Marini, F.; Walczak, B.; Roger, J.-M. *J. Chemom.* **2019**, e3164.

algorithms aimed at signal processing are often integrated with a smoothing step; a typical example being represented by the *Savitzky and Golay algorithm*[258]. Savitzky-Golay smoothing approximates the signals to a polynomial function of appropriate order, calculated within a window of sequential points that is moved across the signal (least-squares fit convolution). The effect of first and second Savitzky-Golay derivatives is shown in Figure A2. The generic polynomial expression[293] on which Savitzky-Golay approach is based is shown in eq. A5:

$$U(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_j x^j = \sum_{q=0}^{j} c_q x^q \qquad \text{(eq. A5)}$$

It has been demonstrated that by applying least-squares-fit convolution to a set of uniformly spaced digital data points, centered around a data point u_i, it is possible to determine a set of n coefficients ($p_s^{(q)}$) that give the smoothed value of the qth derivative according to:

$$\frac{d^q \hat{u}_i}{dx^q} = \sum_{s=-m}^{m} p_s^{(q)} u_{i+s} \qquad \text{(eq. A6)}$$

where $i = m + 1 \div N - m - 1$, with N being the number of points that constitute the spectral window on which to calculate the polynomial. This parameter, together with the degree of the polynomial itself, must be carefully chosen. The disadvantage of this technique is that some data are lost during the fitting, therefore it is not applicable to spectra with few points.

[293] Madden, H.H. *Anal. Chem.* **1978**, *50(9)*, 1383-1386.

Appendix B

B1 Validation

In Chemometrics, empirical models are used. They highlight information coming from real experimental data with the aim of approximating the phenomenon under study and predicting one or more properties of interest. Many models could in principle be calculated on the same set of data, giving rise to a plethora of different results and performances due to the high number of involved factors[294]. The validation is the process of selecting the most appropriate model for the analyzed data and verifying its reliability[295]. In the same way as univariate statistics uses common figures of merit as detection and quantification limits, sensitivity, and so on, the validation of multivariate models requires the definition of suitable diagnostics. This is usually based on the application of an error criteria. The model error, generally estimated with the figures of merit described in Section B2, has to be determined on a dataset different than the one used for calculating the model parameters, in order to avoid overoptimistic conclusions and biased estimates[142]. This is normally accomplished through the use of internal validation (e.g., by means of cross-validation) or of an external test set. The latter (see Figure B1) is meant as a second independent set of data for evaluating the performances and, consequently, calculating the residuals (test set validation). This is the approach which best evaluates the generalizability of the model; therefore, if possible (i.e., if the dataset is composed of a sufficient number of samples), it is the one to be preferred for measuring the predictive ability. On the other hand, cross-validation is based on the repeated resampling of the dataset, into training and test sub-sets, so that at each iteration only a part of the original samples is used for model building while the remaining individuals are left out for validation. Generally, it is rather used for model selection (e.g., fix the optimal complexity) than for the final validation stage because the results can be still biased as the calibration and validation sets are never completely independent on one another.

---

[294] Brereton, R.G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A.; Rodionova, O.; Roger, J.-M.; Walczak, B.; Tauler, R. *Anal. Bioanal. Chem*. **2018**, *410(26)*, 6691-6704.

[295] Harshman, R. In Law, H.G.; Snyder Jr, C.W.; Hattie, J. and Mc Donald R. P. (Eds.) *Research Methods for Multimode Data Analysis* (pp. 566-591), New York, NY: Praeger, **1984**.

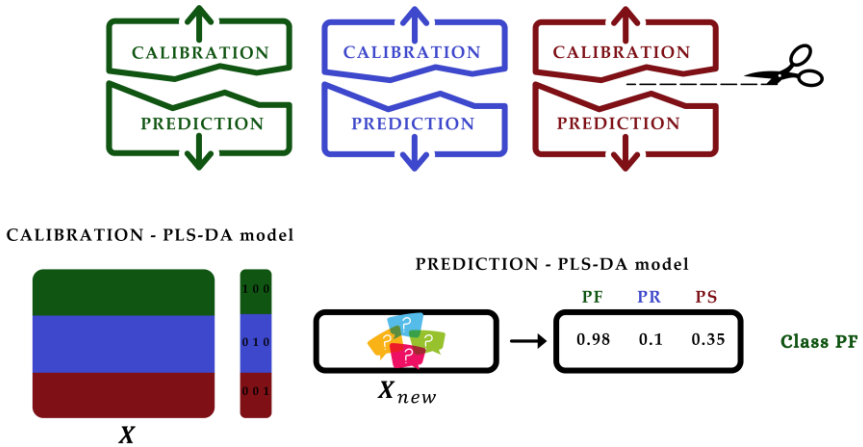


Figure B1. Schematic representation of a data splitting procedure in a classification-context using the Duplex algorithm, applied one class at a time. The example is taken from Section 1.

Hence, a preliminary crucial step in the development of a predictive model is how to perform data splitting, i.e., the choice of the most suitable number and kind of samples (representativity) to insert in the training and test sets to calibrate and adequately validate the model, respectively. Historically, the first algorithm to select a representative subset of the available samples so that they span as uniformly as possible the design space was proposed by Kennard and Stone[296]: given a set of candidate samples, the algorithm aims at selecting the most diverse (i.e., the most distant) ones to be included in the training set. A modification of the algorithm, named Duplex[297], aiming at maintaining a comparable diversity between the two sets has been employed in this thesis.

In the MB analysis performed in this thesis, a non-supervised multi-block method Common components and specific weights analysis[298] (CCSWA), has been used to analyze the datasets prior the application of Duplex splitting on the super-scores matrix derived.

CCSWA is also applied under the acronym *ComDim*[299], which stands for an abbreviation of *Common Dimensions*, giving an idea of the main feature of this method. A detailed description of this algorithm is out of the aim of the discussion; the reader simply takes into account the following assumption. ComDim aims at determining the underlying components together with component-specific block weights (or saliences) which reflect

[296] Kennard, R.W.; Stone, L.A. *Technometrics*. **1969**, *11*, 137-148.
[297] Snee, R.D. *Technometrics*. **1977**, *19*, 415-428.
[298] a) Qannari, E.M.; Wakeling, I.; Courcoux, P.; MacFie, H.J.H. *Food Qual Pref*. **2000**, *11(1-2)*, 151-154; b) Qannari, E.M.; Courcoux, P.; Vigneau, E. *Food Qual Pref*. **2001**, *12(5-7)*, 365-368.
[299] Nielsen, J.P.; Bertrand, D.; Micklander, E.; Courcoux, P.; Munck, L. *J Near Infrared Spec*. **2001**, *9(4)*, 275-285.

the importance that the various blocks of variables attach to these components[300], constrained to be orthogonal.

B2 Figures of merit in validating statistical models

In the regression framework, the most widely used figure of merit is the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\Sigma(\hat{y}-y)^2}{n}} \qquad \text{(eq.B1)}$$

where n is the number of observations in the considered set, y is the real value of the response, and $\hat{y}$ is the response calculated/predicted by the model. If the calibration or the cross-validation or the external set is considered, reference is made to Root Mean Square Error in Calibration (RMSEC), Cross-Validation (RMSECV), or Prediction (RMSEP), respectively.

In the discriminant classification framework[301], several parameters are used to evaluate the goodness of a model; one of these is the so-called *confusion matrix*, i.e., a square matrix of dimensions GxG comparing the experimental with the calculated/predicted classes. For instance, for a two classes classification problem, the *positive* P and the *negative* N, the typical confusion matrix in Table B1 is obtained.

Table B1. Confusion matrix for a two-classes problem.

		Calculated class	
		P	**N**
True class	**P**	TP	FN
	N	FP	TN

TP and TN are the number of P and N samples correctly classified as P and N, respectively; conversely, FP and FN is the number of P and N samples wrongly classified as P and N, respectively. Among the possible parameters derivable from the confusion matrix, the following has been used in this thesis.

[300] Cariou, V.; Qannari, E.M.; Rutledge, D.N.; Vigneau, E. *Food Qual. Prefer*. **2018**, *67*, 27-34.
[301] Ballabio, D.; Grisoni, F.; Todeschini, R. *Chemom. Intell. Lab. Syst*. **2018**, *174*, 33-44.

The Correct Classification Rate ($CCR, \%$) is the estimation of the capability of the model to correctly classify samples. In this thesis, it has been expressed per class, or averaged:

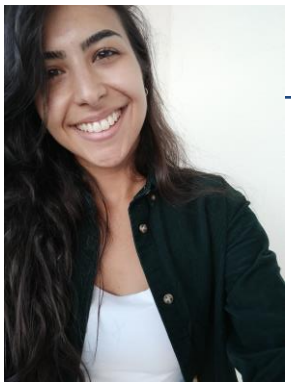
$$\begin{cases} CCR, \%_g = \frac{c_{gg}}{ng} * 100 \\ mean\ CCR, \% = \frac{\sum_{g=1}^{G} CCR, \%_g}{G} \end{cases} \qquad \text{(eq. B2)}$$

where c_{gg} is the diagonal element of the confusion matrix and ng is the number of individuals in the class g.

Another employed index used to assess the classification quality is the accuracy (*Acc,%*), also known as overall classification rate:

$$Acc, \% = \frac{\sum_{g=1}^{G} c_{gg}}{n} * 100 \qquad \text{(eq. B3)}$$

Acc corresponds to the ratio of correct classifications over the total number of samples n in the set and it does not take into account any information on the classification performance of single classes; it is indeed less suitable in the case of unbalanced classes.



Curriculum Vitae et Studiorum

Francesca Di Donato

PhD Candidate in Analytical Chemistry and Chemometrics

https://orcid.org/0000-0003-0633-9016

linkedin.com/in/francesca-di-donato-3b89091b1

Born in Chieti, 14th March 1994
Italian Citizenship
Mobile +39 380 2694048

francesca.didonato3@graduate.univaq.it
didonatofrancesc@gmail.com

Education and Training

November 2018-January 2022

PhD candidate in Università degli Studi dell'Aquila under the supervision of Prof. Angelo Antonio D'Archivio and Dr Alessandra Biancolillo. My research activity has focused on the **development and validation of experimental/statistical approaches for food traceability**.

June 2021-September 2021

I have spent three months as **visiting PhD student** under the supervision of Prof. Rasmus Bro at Københavns Universitet (Denmark), developing skills in the analysis of 3D-data from hyperspectral imaging systems.

2018

Master's Degree in Chemistry, 110/110 cum laude

Internship under the supervision of Prof. Angelo Antonio D'Archivio, Università degli Studi dell'Aquila.

Thesis title: Detection of Plant-Derived Adulterants in Saffron (Crocus sativus L.) by HS-SPME/GC-MS Profiling of Volatiles and Chemometrics

---

2016

Bachelor's Degree in Chemistry and Material Science, 110/110 cum laude

Internship under the supervision of Prof. Fabrizio Ruggieri, Università degli Studi dell'Aquila.

Internship-Thesis title: UHPLC Analysis of Saffron (Crocus sativus L.): Optimization of Separation Using Chemometrics and Detection of Minor Crocetin Esters

Meetings attended

Conferences

2021

XXVII Congresso Nazionale della Società Chimica Italiana, Milano, Italy

Webinar NIR spectral imaging, Dipartimento di Farmacia, Genoa, Italy

TMS Day of Science, Webex Conference

ROAD TO CAC2022, Zoom Conference

I giovani e la chimica in Abruzzo, Webex Conference

NIR Italia online 2021, Zoom Conference

2020

Online day of discussion on Experimental Design

Webinar Buone pratiche di analisi termica (Mettler Toledo)

Webinar Buone pratiche di spettrofotometria UV-VIS (Mettler Toledo)

2019

XXVIII Congress of the Analytical Chemistry Division, Bari, Italy

XLVIII National Congress on Magnetic Resonance, L'Aquila, Italy

La cromatografia liquida ad alta prestazione (HPLC) nell'analisi e preparazione di sostanze d'interesse biologico, L'Aquila, Italy

Workshop Gruppo di Chemiometria, Bergamo, Italy

2018

XXXVI Convegno Interregionale TUMA 2018, Pisa, Italy

Tecniche Chemiometriche e Strumentali di supporto all'Analisi Sensoriale e all'Agricoltura di Precisione, Salerno, Italy

New Trends in Liquid Chromatography and Sample Preparation, Chieti, Italy

Oral presentations

2021

XXVII Congresso Nazionale della Società Chimica Italiana, Milano, Italy. *Authentication of donkey milk by Near Infrared Spectroscopy coupled with chemometric classifiers*

TMS Day of Science (selected oral presentation), L'Aquila, Italy. *HS-SPME/GC-MS volatile fraction determination and Chemometrics for the discrimination of typical Italian Pecorino cheeses*

I giovani e la chimica in Abruzzo, L'Aquila, Italy. *Geographical discrimination of high value Italian Bell Pepper (Capsicum annuum) varieties by means of ICP-OES multi-elemental analysis coupled with Chemometrics*

NIR Italia online 2021. Sequential Data Fusion *Techniques for the authentication of the P.G.I. Senise ("Crusco") pepper bell*. https://doi.org/10.5281/zenodo.455054

2019

Workshop Gruppo di Chemiometria, Bergamo, Italy. *Discrimination of Potato (Solanum tuberosum L.) Accessions Collected in Majella National Park (Abruzzo, Italy) Using Mid-Infrared Spectroscopy and Chemometrics Combined with Morphological and Molecular Analysis*

2018

XXXVI Convegno Interregionale TUMA 2018, Pisa, Italy. *Analisi SPME-HS/GC-MS del profilo dei volatili per la determinazione dei principali adulteranti di origine vegetale nello zafferano*

Poster presentations

2021

TMS Day of Science, L'Aquila, Italy. *HS-SPME/GC-MS volatile fraction determination and Chemometrics for the discrimination of typical Italian Pecorino cheeses*

2019

XXVIII Congress of the Analytical Chemistry Division, Bari, Italy. *Characterization and discrimination of typical Italian Pecorino cheeses by means of HS-SPME/GC-MS aroma profiling combined with Chemometrics*

XLVIII National Congress on Magnetic Resonance, L'Aquila, Italy. *Multi-methodological NMR-DSC-IR approach for the characterization of typical Pecorino cheeses*

2018

XXVII Congress of the Analytical Chemistry Division, Bologna, Italy. *Determination of plant-derived contaminants in saffron by means of SPME-HS/GC-MS aroma profiling combined with Chemometrics*

Congress on "New Trends in Liquid Chromatography and Sample Preparation", Chieti, Italy. *Development of a UHPLC method for the separation of crocins in saffron extracts using Artificial Neural Networks*

Formation

Doctoral schools

2021

PhD School of SCIENCE, University of Copenhagen (online classes). **International School of Chemometrics 2021**:

MATLAB-Introduction to Matlab for Multivariate Data Analysis,

BASIC-Basic introduction to chemometrics,

INTERMEDIATE-Intermediate topics on chemometrics,

CHALLENGES-Topics for a further understanding of advanced modelling methods,

MCR-Multivariate Curve Resolution,

MSPC-Multivariate Statistic Process Control

2019

Didattica e ricerca educativa. Division of Didattica Chimica-SCI. San Miniato, Italy

University of Genoa, Italy. **Multivariate Analysis**:

PCA and N-way PCA,

Classification and Class Modelling Techniques,

Multivariate Calibration Techniques

University of Genoa, Italy. **Experimental Design**:

Full Factorial Designs and Fractional Factorial Designs,

Screening Designs,

Response Surface Methodology (Central Composite Design and Doehlert Design) and Multicriteria Decision Making,

D-Optimal Designs,

Mixture Designs and Mixture-Process Designs

Courses

2021

Seminari Proprietà Intellettuale e Industriale–University of L'Aquila, L'Aquila, Italy

---

2019

Data Analytics and Data Driven Decision. Department of Information Engineering, Computer Science and Mathematics. University of L'Aquila, L'Aquila, Italy

Applied Chemometrics with MATLAB. Department of Chemical Sciences. La Sapienza, Rome, Italy

Academic English for Doctoral Students (B2 level). University of L'Aquila, L'Aquila, Italy

---

2018

24 Crediti Formativi Universitari (CFU). University of L'Aquila, L'Aquila, Italy

Funds and awards

September 2021.

Fellowship by the Analytical Chemistry Division to attend with an oral presentation the XXVII Congresso Nazionale SCI

July 2021.

Best Elevator Pitch award at TMS Day of Science 2021 from courtesy of AUROGENE s.r.l.

November 2018.

3-year PhD **scholarship** in "Physical and Chemical Sciences" from University of L'Aquila

Academic Experiences

January 2021-October 2021.

TMS (Thursday Morning Science) **Committee member** (Social Media group). TMS is a group of researchers of the University of L'Aquila involved in the organization of scientific talks and conferences in English.

November 2018-October 2021.

Training of student interns in the Analytical research group

Tutoring and supporting bachelor students during the "Analytical laboratory" held by Prof. Angelo Antonio D'Archivio, Prof. Fabrizio Ruggieri and Dr Alessandra Biancolillo

Publications

Submitted Papers

Marchi, L.; Krylov, I.; Roginski, R.T.; Wise, B.; **Di Donato, F.**; Nieto-Ortega, S.; Pereira, J.F.Q.; Bro, R. *Automatic hierarchical model builder. Submitted to Journal of Chemometrics.*

Di Donato, F.; Biancolillo, A.; Ferretti, A.; D'Archivio, A.A.; Marini, F. *Near Infrared Spectroscopy coupled to chemometrics for the authentication of donkey milk. Submitted to Food Control.*

Di Donato, F.; Biancolillo, A.; Foschi, M.; D'Archivio, A.A. *Application of SPORT algorithm on ATR-FTIR data: a rapid and green tool for the characterization and discrimination of three typical Italian Pecorino cheeses. Submitted to Journal of Food Composition and Analysis.*

Di Donato, F.; Biancolillo, A.; Foschi, M.; Di Cecco, V.; Di Martino, L.; D'Archivio A.A. *Authentication of typical Italian bell pepper spices by ICP-OES multi-elemental analysis combined with SIMCA class modelling. Accepted pending revisions for Food Chemistry*.

Published Papers

Nori, V.; Sinibaldi, A.; Giorgianni, G.; Pesciaioli, F.; **Di Donato, F.**; Biancolillo, A.; Landa, A.; Carlone, A. *DoE-driven development of an organocatalytic enantioselective addition of acetaldehyde to nitrostyrenes in water. Chemistry-A European Journal,* **2022**, e202104524. https://doi.org/10.1002/chem.202104524

Biancolillo, A.; Foschi, M.; Di Micco, M.; **Di Donato, F.**; D'Archivio, A.A. *ATR-FTIR-based rapid solution for the discrimination of lentils from different origins, with a special focus on PGI and Slow Food typical varieties. Microchemical Journal,* **2022**, *178(1)*, 107327. https://doi.org/10.1016/j.microc.2022.107327

Di Donato, F.; Foschi, M.; Vlad, N.; Biancolillo, A.; D'Archivio, A.A. *Multi-elemental composition data handled by Chemometrics for the discrimination of high-value Italian Pecorino Cheeses. Molecules,* **2021**, *26(22)*, 6875. https://doi.org/10.3390/molecules26226875

Di Donato, F.; Squeo, F.; Biancolillo, A.; Rossi, L.; D'Archivio, A.A. *Characterization of high value Italian chickpeas (Cicer arietinum L.) by means of ICP-OES multi-elemental analysis coupled with chemometrics. Food Control*, **2021**, *131(9)*, 108451. https://doi.org/10.1016/j.foodcont.2021.108451

Reale, S.; Di Cecco, V.; **Di Donato, F.**; Di Martino, L.; Manzi, A.; Di Santo, M.; D'Archivio A.A. *Characterization of the Volatile Profile of Cultivated and Wild-Type*

Italian Celery (Apium graveolens L.) Varieties by HS-SPME/GC-MS. Applied Sciences, **2021**, *11(13)*, 5855. https://doi.org/10.3390/app11135855

Di Donato, F.; Biancolillo, A.; Mazzulli, D.; Rossi, L.; D'Archivio A.A. *HS-SPME/GC–MS volatile fraction determination and chemometrics for the discrimination of typical Italian Pecorino cheeses. Microchemical Journal*, **2021**, *165*, 106133. https://doi.org/10.1016/j.microc.2021.106133

Aït-Kaddour, A.; Hassoun, A.; Bord, C.; Schmidt-Filgueras, R.; Biancolillo, A.; **Di Donato, F.**; Tümay Temiz, H.; Cozzolino, D. *Application of Spectroscopic Techniques to Evaluate Heat Treatments in Milk and Dairy Products: an Overview of the Last Decade. Food and Bioprocess Technology*, **2021**, *14*, 781-803. https://doi.org/10.1007/s11947-021-02607-0

Biancolillo, A.; **Di Donato, F.**; Merola, F.; Marini, F.; D'Archivio A.A. *Sequential Data Fusion Techniques for the Authentication of the P.G.I. Senise ("Crusco") Bell Pepper. Applied Sciences*, **2021**, *11(4)*, 1709. https://doi.org/10.3390/app11041709

Di Donato, F.; Gornati, G.; Biancolillo, A.; D'Archivio, A.A. *ICP-OES analysis coupled with chemometrics for the characterization and the discrimination of high added value Italian Emmer samples. Journal of Food Composition and Analysis*, **2021**, *98*, 103842. https://doi.org/10.1016/j.jfca.2021.103842

Di Donato, F.; D'Archivio, A.A.; Maggi, M.A.; Rossi, L. *Detection of Plant-Derived Adulterants in Saffron (Crocus sativus L.) by HS-SPME/GC-MS Profiling of Volatiles and Chemometrics. Food Analytical Methods*, **2021**, *14*, 784-796. https://doi.org/10.1007/s12161-020-01941-x

Hassoun, A.; Shumilina, E.; **Di Donato, F.**; Foschi, M.; Simal-Gandara, J.; Biancolillo, A. *Emerging Techniques for Differentiation of Fresh and Frozen–Thawed Seafoods: Highlighting the Potentialof Spectroscopic Techniques. Molecules*, **2020**, *25(19)*, 4472. https://doi.org/10.3390/molecules25194472

Hassoun, A.; Guðjónsdóttir, M.; Prieto, M.A.; Garcia-Oliveira, P.; Simal-Gandara, J.; Marini, F.; **Di Donato, F.**; D'Archivio, A.A.; Biancolillo, A.A. *Application of Novel Techniques for Monitoring Quality Changes in Meat and Fish Products during Traditional Processing Processes: Reconciling Novelty and Tradition. Processes*, **2020**, *8(8)*, 988. https://doi.org/10.3390/pr8080988

Di Donato, F.; Di Cecco, V.; Torricelli, R.; D'Archivio, A.A.; Di Santo, M.; Albertini, E.; Veronesi, F.; Garramone, R.; Aversano, R.; Marcantonio, G.; Di Martino, L. *Discrimination of Potato (Solanum tuberosum L.) Accessions Collected in Majella National Park (Abruzzo, Italy) Using Mid-Infrared Spectroscopy and Chemometrics Combined with Morphological and Molecular Analysis. Applied Sciences*, **2020**, *10(5)*, 1630. https://doi.org/10.3390/app10051630

Iannella, M.; Console, G.; D'Alessandro, P.; Cerasoli, P.; Mantoni, C.; Ruggieri, F.; **Di Donato, F.**; Biondi, M. *Preliminary Analysis of the Diet of Triturus carnifex and Pollution in Mountain Karst Ponds in Central Apennines. Water*, **2020**, *12(1)*, 44. https://doi.org/10.3390/w12010044

D'Archivio, A.A.; **Di Donato, F.**; Foschi, M.; Maggi, M.A.; Ruggieri, F. *UHPLC Analysis of Saffron (Crocus sativus L.): Optimization of Separation Using Chemometrics and Detection of Minor Crocetin Esters. Molecules*, **2018**, *23(8)*, 1851. https://doi.org/10.3390/molecules23081851

L'Aquila, May 2022 Francesca Di Donato