UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Scienze Chimiche e Geologiche

Dottorato di ricerca in
Models and methods for material and environmental sciences

Ciclo XXXVI

Chemometric aided quality assessment from lab to plant in an Industry 4.0 context

Candidate
Alessandro D'Alessandro

Tutor
Prof. Marina Cocchi

Co-tutor
Prof. Caterina Durante

Course Coordinator
Prof. Stefano Lugli

To Lidia, Paolo, and Simona

"The illiterate of the 21st century will not be those who cannot read or write, but those who cannot learn, unlearn, and relearn."

Alvin Toffler (writer, businessman, futurist, 1928-2016)
(from the book «Future Shock», 1970)

"Do. Or do not. There is no try."

Yoda
(from the movie «The Empire Strikes Back», 1980)

ABSTRACT

English

Product quality is a "must" for every producer. For a food company like Barilla, where I work, this is very relevant because food is strongly linked to our emotions, our health, and our well-being. Regarding this, in fact, the Company's 'Mission' is: "*The joy of food for a better life. Bringing people closer to the joy of good food and making quality the choice for a better life, from each individual to the planet"*. This explains why there is a strong commitment to every quality-related issue in the Company.

One of the tasks of the Analytical Food Science Research and Development Laboratory, where I work, is to develop new methods and tools to assess the quality of our products, both in a research and industrial context. The use of chemometrics in my work has grown over time because of its great ability to extract information from large amounts of data and the ability to present this information concisely and effectively. In some cases, the use of chemometric techniques is essential and it is not possible to analyse the data in any other way. Whitin an industrial context that is rapidly moving toward an Industry 4.0 context, more and more data are being produced from all the sensors installed in production lines, data that need to be analysed real-time and evaluated in the appropriate way.

"*Pesto Genovese*" is an Italian green sauce made mainly of basil and olive oil, cheese, pine nuts and garlic, has a unique flavour known and appreciated all over the world. In this Thesis project, Barilla's production of "*Pesto alla Genovese*" was used as a benchmark.

The objective of the Thesis project was to develop analytical-chemometric methods suitable for evaluating (i) the characteristics of the main raw material, basil, and of the finished product, pesto, in the most rapid and effective way, in a quality laboratory analysis context; and, (ii) the characteristics of the raw material, the production intermediate and the finished product in order to develop models for real-time quality monitoring, in a process monitoring context.

From the analytical point of view, approaches based on rapid, non-destructive techniques have been developed, such as electron nose (based on gas chromatography), near-infrared (NIR) spectroscopy in its various implementations including multi- and hyper-spectral imaging. Chemometric approaches, which are essential for efficiently extracting the information obtained through these techniques, have ranged from exploratory multivariate analysis, multivariate variance analysis methods, image analysis methods, to the development of multivariate control charts and predictive models, always evaluating appropriate pre- and post-processing methods.

The work done has demonstrated, through several real cases, how chemometrics is an indispensable support for obtaining information that would otherwise not be accessible and can provide powerful tools for real-time control of critical raw materials, process, and product.

Despite the specific topic related to *pesto*, the approaches developed are general and extensible to other products/processes in the food industry. The main challenge was to transfer the methodological know-how to this application context.

In conclusion, the original idea of this industrial PhD project to build a "statistical tool" for my daily work was successfully realized. In addition, the cases studied in the production environment open potential new applications with a strong impact on improving the possibility of process control and designed quality.

Italiano

La qualità del prodotto è un “must” per ogni produttore. Per un'azienda alimentare come la Barilla, dove lavoro, ciò è molto rilevante perché il cibo è fortemente legato alle nostre emozioni, alla nostra salute e al nostro benessere. Riguardo a ciò, infatti, la ‘Mission’ dell'Azienda è: “La gioia del cibo per una vita migliore. Avvicinare le persone al piacere del buon cibo e fare della qualità la scelta per una vita migliore, di ogni individuo e del pianeta”. Questo spiega perché in Azienda è presente un forte impegno verso ogni tema legato alla qualità.

Uno dei compiti del Laboratorio di Ricerca e Sviluppo di Scienze Alimentari Analitiche, dove lavoro, è quello di sviluppare nuovi metodi e strumenti per valutare la qualità dei nostri prodotti, sia in un contesto di ricerca che in quello industriale. L'uso della chemiometria nel mio lavoro è cresciuto nel tempo grazie alla sua grande capacità di estrarre informazioni da grandi quantità di dati e alla possibilità di presentare queste informazioni in modo sintetico ed efficace. In alcuni casi, l'uso di tecniche chemiometriche è essenziale e non è possibile analizzare i dati in altro modo. In particolare, in un mondo industriale che si muove rapidamente verso un contesto Industria 4.0, vengono prodotti sempre più dati da tutti i sensori installati nelle linee di produzione, dati che necessitano di essere analizzati real-time e valutati nel modo appropriato.

Il “Pesto Genovese” è una salsa verde italiana a base principalmente di basilico, olio d'oliva, formaggio, pinoli e aglio e ha un sapore unico conosciuto ed apprezzato in tutto il mondo. In questo progetto di Tesi, la produzione di “*Pesto alla Genovese*” Barilla è stata utilizzata come benchmark.

Obiettivo del progetto di tesi è stato quello di sviluppare dei metodi analitici-chemiometrici adeguati a valutare: (i) le caratteristiche della principale materia prima, il basilico, e del prodotto finito, il pesto, nel modo più rapido ed efficace, in un contesto di analisi laboratorio qualità; (ii) le caratteristiche della materia prima, dell’intermedio di produzione e del prodotto finito allo scopo di sviluppare modelli per il monitoraggio real-time della qualità, in un contesto di monitoraggio di processo.

Dal punto di vista analitico sono stati sviluppati approcci basati su tecniche rapide e non-distruttive, quali il naso elettronico (basato sulla gas-cromatografia), la spettroscopia nel vicino infrarosso (NIR) nelle sue diverse implementazioni incluso l’imaging multi e iper-spettrale. Gli approcci chemiometrici, fondamentali per estrarre in modo efficiente le informazioni ottenute attraverso queste tecniche, hanno spaziato dall’analisi multivariata esplorativa, metodi di analisi di varianza multivariata, metodi di analisi di immagini, allo sviluppo di carte di controllo multivariate e modelli predittivi, sempre valutando gli opportuni metodi di pre- e post- processing.

Il lavoro svolto ha dimostrato, attraverso diversi casi reali, come la chemiometria sia un supporto indispensabile per ottenere informazioni che altrimenti non sarebbero accessibili e possa fornire potenti strumenti per il controllo in tempo reale delle materie prime critiche, del processo e del prodotto.

Nonostante il tema specifico relativo al pesto, gli approcci sviluppati sono generali ed estensibili ad altri prodotti/processi dell’industria alimentare. La sfida principale è stata trasferire in questo contesto applicativo il know-how metodologico.

In conclusione, l’idea originale di questo progetto di Dottorato industriale di costruire uno “strumento statistico” per il mio lavoro quotidiano è stata realizzata con successo. Inoltre, i casi studiati in ambito produttivo, aprono a nuove potenziali applicazioni con un forte impatto sul miglioramento della possibilità di controllo del processo e della qualità progettata.

LIST OF PUBLICATIONS

Primary works

Published

[1] Characterization of Basil Volatile Fraction and Study of Its Agronomic Variation by ASCA
D'Alessandro, A.; Ballestrieri, D.; Strani, L.; Cocchi, M.; Durante, C.
Molecules 2021, 26, 3842.
https://doi.org/10.3390/molecules26133842

[2] Fast GC E-Nose and Chemometrics for the Rapid Assessment of Basil Aroma
Strani, L.; D'Alessandro, A.; Ballestrieri, D.; Durante, C.; Cocchi, M.
Chemosensors 2022, 10, 105.
https://doi.org/10.3390/chemosensors10030105

[3] A Feasibility Study towards the On-Line Quality Assessment of Pesto Sauce Production by NIR and Chemometrics
Tanzilli, D.; D'Alessandro, A.; Tamelli, S.; Durante, C.; Cocchi, M.; Strani, L.
Foods 2023, 12, 1679.
https://doi.org/10.3390/foods12081679

Under submission

[4] *Pesto alla genovese* hyperspectral images interpreted by chemometrics methodologies: a case study
D'Alessandro, A.*; Strani, L.; Tanzilli, D.; Mas, S.; Ryckewaert, M.; Roger, J.M.; Cocchi, M.
Food Chemistry

In preparation

[5] A Comparative Study of Chemometrics and Deep Learning on Semantic Segmentation Classification
Tanzilli, D.; D'Alessandro, A; Løve Hinrich, J.; Amigo, J.; Cocchi M.

[6] Multispectral image analysis for the characterisation of basil leaves.
D'Alessandro, A.; Cocchi, M.

Auxiliary works

Published

[7] Near Infrared and UV-Visible spectroscopy coupled with chemometrics for the Characterization of Flours from Different Starch Origins
Pellacani, S.; Borsari, M.; Cocchi, M.; D'Alessandro, A.; Durante, C.; Farioli, G.; Strani, L.
Chemosensors 2024, 12, 1.
https://doi.org/10.3390/chemosensors12010001

[8] How Chemometrics Can Fight Milk Adulteration
Grassi, S.; Tarapoulouzi, M.; D'Alessandro, A.; Agriopoulou, S.; Strani, L.; Varzakas, T.
Foods 2023, 12, 139.
https://doi.org/10.3390/foods12010139

TABLE OF CONTENTS

1INTRODUCTION

1.1 Context

The basic aim in food industry research and development (R&D) is to create new products and launch them successfully on the market. More specific aims, which are strategic for R&D, include: offering a wider choice of food items to the consumers; enhancing good sensory perception that makes food more appealing; improving nutritional value to meet dietary needs; improving food safety; adding convenience; and finally, reducing production costs which may allow product prices to be lowered. These beneficial outcomes can be reached either from constant gradual product improvement or by introducing a significant product change in a single step. The latter situation usually takes place when a new technology - crop, ingredient, process, storage – is introduced, as well as when a new understanding of consumer needs is achieved.

R&D covers the total food system chain and needs multidisciplinary research because the scientific base of the food system takes roots in diverse disciplines such as chemical, biological sciences and food technologies. The objective is always consumer satisfaction, but in a broad meaning that includes not only the individual perception but also consumer health and safety together with the environmental sustainability and the well-being of those who produce the food.

Consumer satisfaction is strictly linked to the products quality. So, in food industrial production, guaranteeing a constant quality of the final product is a must, especially for brands with a high reputation. A great effort is deployed to design processes robust enough to always ensure the desired quality, compensating for the “physiological” variability of food raw materials and processes. The concept behind this is the Process Analytical Technology (PAT) [1,2,3,4,5,6] linked to the Quality by Design (QbD) [7] paradigm, which is based on the concept that the quality of the (food) products can and should be ensured by process design and control (i.e. integrated into the process) and not only provided by post-production quality testing. Of course, this does not eliminate the need to apply quality control protocols to continuously monitor the final product as well as the process itself [8,9].

To set up a PAT-QbD framework two basic requirements must be met: i) the implementation of on-line sensors and ii) the use of multivariate data analysis tools to extract, integrate and utilise the information provided by process and analytical sensors and link it to the product quality assessment. This framework will allow reaching process knowledge, such as what is the natural process variability, what are the most critical factors to control, how to implement a process monitoring/control system [10].

This reflects in food scientists and technologists facing, during the last 30 years, increasing massive amounts of data derived from the use of different measuring devices (e.g., instrumental, and sensory data), the integration of different analytical techniques and processes during the analysis and production of foods. Therefore, complementary disciplines and tools, such as statistics and chemometrics, experimental design (DOE), Multivariate data analysis (MVDA), multivariate statistical process control (MSPC), [11,12,13] add to the more traditional ones used in food science, and they have become essential in modern sciences and are an integral component in the day-to-day foods analysis.

It became so clear how much is felt the need to have appropriate methods to characterize row materials as well as production intermediates and final products. Appropriate analytical methods should be fast, non-destructive and, possibly, easy-to-use, considering their use in industrial context, and unavoidably they should be supported by data analysis tools.

In this context, the main aim of my thesis project was to assemble a “*toolbox*” of knowledge and chemometric techniques allowing me developing proper analytical methods in my Company, in both R&D and industrial contexts.

To do that, some cases of study related to the production chain of green sauce “*Pesto alla Genovese*” were selected as benchmarks for applying chemometric tools and improve data

analysis strategies. Of course, the approaches used are extendable to any other products or production plant.

"*Pesto alla Genovese*" [14] is a green sauce inspired to the "PESTO GENOVESE" [15,16] name that associates to the original recipe of Italian traditional Basil Pesto sauce done with the seven ingredients contemplated by the Consortium for protection and guarantee of the ancient regional heritage, which are: PDO Genovese basil, Extra virgin olive oil, PDO Parmigiano Reggiano (as well allowed Grana Padano variant), PDO Pecorino Sardo, Pine nuts, Garlic and Salt. It has a unique flavour known and appreciated all over the world. Hereafter we will refer always to "*Pesto alla Genovese*".

The industrial production of "*Pesto alla Genovese*" requires the accurate control of the raw materials quality. Basil is one of the main ingredients of pesto sauce, in terms of importance and quantity. Its evaluation is nowadays still done visually inspecting a small part of the huge quantity of incoming basil in the production plant. This could be a weakness considering how its characteristics like aroma, plant colour, leaves to stems ratio, and defects are subject to variability while being so relevant for the final product quality.

Analogously for the final product "*Pesto alla Genovese*" the aroma (in large part influenced by the basil) is one of the key quality factors, together with the physical structure related to the product creaminess and colour.

Therefore, there is a need to select appropriate tools, analytical and statistical, to proper characterize basil and pesto. Moreover, in an industrial contest, to minimize the number of routine quality analyses, it is also important together with assessing which analysis describes at best the product quality.

1.2 State of the art

The concept of *process analytics* (PA) was probably born since 1940s in Germany in chemical and petrochemical industries. In these industries the PA was implemented as chemical or physical analysis [17] of materials carried out during the process. In the following twenty years it was also adopted by nuclear power plants [18]. The concept of Quality by Design was [19] first proposed by Joseph M. Juran in 1992 in some publications, mainly in Juran Quality by Design [20]. The basic idea is that quality can be planned. It was primarily [21] adopted by the automotive industry and then the US Food and Drug Administration (FDA) used it for the process of drug discovery, development, and manufacturing in early 2000 [22], introducing the concept of Process Analytical Technology (PAT).

In food industry the adoption of QbD had have a slower speed [23], probably due to the relevant difference respect to pharma industries in terms of profit margins and consequently on invested money in more sophisticated technologies.

In a recent paper [24] the implementation of the QdB/PAT tools in food industry has been studied. Results indicates that *"QdB/PAT bases and tool are still rarely implemented in food industry".* There could be many causes for this, including the preference of the companies to evaluate the quality of products with a more "classical" off-line analysis using laboratory-based analytical methods [25,26]. So, in the Perez-Beltran paper [5 cit.] just 23 studies of QbD/PAT application in food industry were found, and this although QbD/PAT tools have been demonstrated their huge impact in improving process understanding and control and saving money by reducing the number of non-compliant products that have to be discarded.

In food context the definition of quality includes more than one criterion: authenticity (food authentic, traditional, or natural and not adulterated during production, processing, or storage) sometimes also expressed as "integrity"; function (i.e. cooks well); biological activity (positive or negative interaction with body's functions); nutrition (contribution to a healthy diet); sensorial experience (smell, taste, texture) and ethical (environmental, social, and ethical aspects).

Application of Quality by Design requires a change from the classical inferential monitoring and controls of simple parameters in production (pH, temperature, pressure), most often done one parameter a time, towards core parameters that requires real-time measurements during the production process, by on-line or in-line techniques followed by multivariate data analysis [27,28]

to consider correlation structure of the different parameters. Moreover, the advent of the Industry 4.0, the so-called fourth industrial revolution, will open new scenarios. The term “Food processing 4.0” has been proposed [24] to indicate the industrial revolution 4.0 also in the food production. The Food processing 4.0 concept denotes processing food in a high technologized environment in which more attention will be posed, not only to the classical quality parameter (already cited), but also to environmental impact of the production processing in terms of consumption of water, energy, wastes, etc.

The application of Food Processing 4.0 requires, from one side a new production environment with an increased level of interconnections (sensors, devices, measurement systems, machinery, data storage), and from the other side a fundamental aspect of interdisciplinary in chemical, physical, digital, and biological fields [29]. Despite this huge re-conversion required and the fact that food industry has typically less money to invest compared to pharma or biotech, in the last ten years the interest in this topic is exponential increasing. In parallel, it became more and more necessary to dispose of appropriate analytical techniques and mathematical tools to be applied.

In another study Djekic et al. [30], conducted a survey of more than 200 European industries and they found that even if they implemented some QbD approach, their applications consisted of rather simplified models that did not evaluate all the QbD aspects (i.e. safety conditions or environmental impact). Further, it was pointed out that the application of mathematical models in food companies has not yet been a matter of interest. The study also identified some reasons to explain the absence of multivariate tools in the food companies: limited background knowledge on modelling; software that is not user-friendly; instability of processes when introducing experimental tests; additional cost of new experiments to be planned in the initial model building phase (although money is saved in the long term and the balance would be favourable, there is little awareness of it); high confidentiality of the studies already carried out, which hinders the free publication of the results in scientific journals.

The conclusion is that it is necessary to start spreading the QbD/PAT approach in a broader and more complete mode in the food industry context.

This could be done faster and better by improving the cooperation between Academia and companies and a PhD Thesis like mine is a promising first step.

1.3 Thesis aims and outlines

During the PhD Thesis project different analytical methods and modelling strategies were applied to evaluate the characteristics of basil and pesto in the fastest and most effective way. Despite the specific benchmark, the developed approaches are general and extendable to other product/processes in the food industry.

The need for proper analytical methodologies embraces two main areas: the R&D area in which the characterization and evaluation of new basil chemotypes or new pesto prototypes has been exploited, and the Production area, where the focus has mainly been on controlling the homogeneity of the production in time (possibly real-time).

In both cases chemometrics is fundamental to efficiently extract proper information from analytical data.

In the scheme below (Figure 1-1) is shown a synthesis of the work undertaken during the three years.

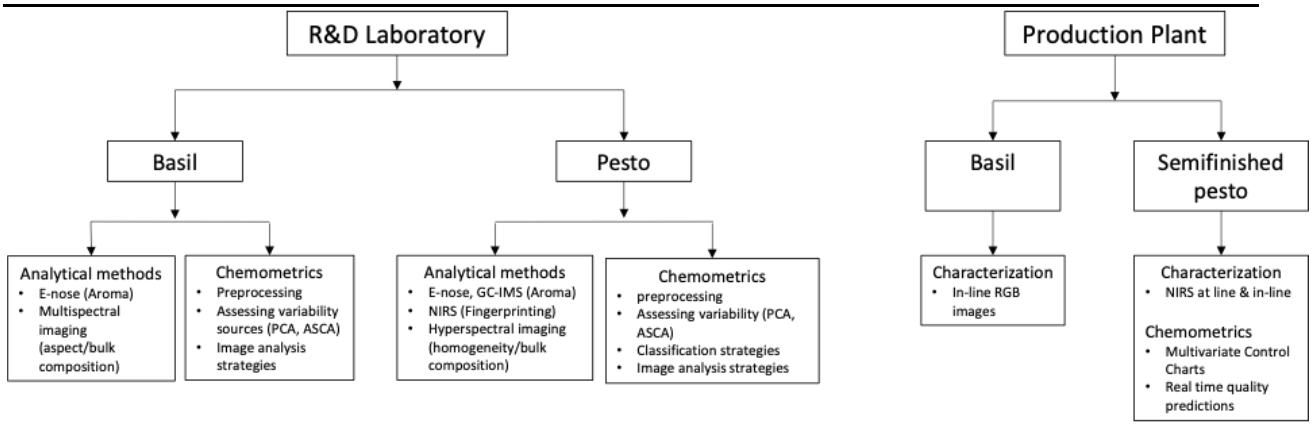


Figure 1-1. Summary of the global pathway

The Thesis organization follows from this scheme; below a brief description of the content of the different chapters is illustrated.

In the studies reported in Chapters 4 the focus has been on aroma characterisation of both basil and pesto since flavour is a very important food quality attribute for consumers. Several analytical techniques have been evaluated coupled with the application of proper data elaboration strategies and tools. In particular, two analytical techniques, such as Head Space Fast Gas Chromatography electronic-nose (HS-FGC-e-nose) and Head Space Gas Chromatography-Ion Mobility Spectrometry (HS-GC-IMS), were esteemed promising and applied to evaluate basil chemotypes in agronomical studies with the main aim of selecting the basil chemotypes holding the best aroma profile. The multivariate data elaboration was essential in both cases.

In the studies reported in Chapter 5, several analytical techniques have been evaluated to assess the most effective one for discriminating pesto samples obtained with different basil types. In this case, data curation, pre-processing, and exploration prior to classification were fundamental. Moreover, data fusion at low level permitted to better understand which of the evaluated techniques were more effective.

Chapter 6 was dedicated to imaging methodologies applied to both basil and pesto. Different imaging system were evaluated, from the simpler and common RGB imaging to hyperspectral imaging systems in the Vis and Near Infrared ranges. These techniques require a pool of chemometric techniques and image analysis tools to extract the diverse and relevant information aiming at interpretable results.

Finally, Chapter 7 was dedicated to a feasibility study for real-time on-line quality assessment in the production plant. Here, chemometric tools for Multivariate Statistical Process Monitoring (MSPC) and predictive modelling were applied. The main practical issues to be faced were exploited and discussed. While, the on-line monitoring system needs to be improved, it has been possible to demonstrate, as proof of concept, the possibility to predict in advance final product chemical parameters from NIR on-line on a semi-finished basis.

1 Guidance for Industry, PAT – A Framework for Innovative Pharmaceutical Manufacturing and Quality Assurance, Draft Guidance" available form www.fda.gov/cder/OPS/PAT.htm

2 Davies T. "What is PAT?" Spectroscopy Europe April/May 2004; 16(2) 33-34.

3 Juran JM. "The Quality Trilogy: A Universal Approach to Managing for Quality". Quality Progress. 1986; 19(8): 19-24.

4 Callis JB, Illman DL and Kowalski BR. “Process Analytical Chemistry” Analytical Chemistry. 1987; 59:624A-637

5 Pérez-Beltrán CH, Jiménez-Carvelo AM, Torrente-López A et al. QbD/PAT—State of the Art of Multivariate Methodologies in Food and Food-Related Biotech Industries. Food Eng Rev. 2023;15, 24–40. https://doi.org/10.1007/s12393-022-09324-0

6 Davis JR and Wasynczuk J. “The Four Steps of PAT Implementation” Pharmaceutical Engineering. January/February 2005; 10-22

7 van den Berg FWJ “Optimal Process Analyzer Selection and Positioning for Plant-Wide Monitoring” Ph.D. Thesis University of Amsterdam (2001) available from www.models.kvl.dk https://sid.erda.dk/share_redirect/d9hH4DvvY1

8 Smilde AK, van den Berg FWJ and Hoefsloot HCJ “How to choose the right process analyzer” Analytical Chemistry. 2002; 74/13:368A-373

9 Nychas GJE, Panagou EZ, Mohareb F. “Novel approaches for food safety management and communication” Current Opinion in Food Science. 2016; 12:13-20. https://doi.org/10.1016/j.cofs.2016.06.005

10 Kourti TH and MacGregor J. “Tutorial: Process Analysis, monitoring and diagnosis, using multivariate projection methods” Chemometrics and Intelligent Laboratory Systems. 1995; 28:3-21

11 Buvé C, Saeys W, Rasmussen MA, Neckebroeck B, Hendrickx M Grauwet et al. Application of multivariate data analysis for food quality investigations: An example-based review, Food Research International.2022; 151: 110878. https://doi.org/10.1016/j.foodres.2021.110878.

12 Zhang Z, Li Y, Zhao S, Qie M, Bai L Gao Z et al. Rapid analysis technologies with chemometrics for food authenticity field: A review, Current Research in Food Science.2024;8:100676. https://doi.org/10.1016/j.crfs.2024.100676.

13 Schweitzer M et al.Implications and Opportunities of Applying QbD Principles to Analytical Measurements. Pharmaceutical Technology. 2010; 34 (2): 52–59.

14 https://www.barilla.com/it-it/prodotti/sughi/pesto-alla-genovese [Accessed January 2024]

15 https://www.mangiareinliguria.it/pesto-genovese/consorzio-pesto-genovese [Accessed January 2024]

16 https://www.mangiareinliguria.it/pesto-genovese/ricetta-pesto-genovese [Accessed January 2024]

17 Festing MFW. Principles: The need for better experimental design, Trends in Pharmacological Sciences, 2003; 24(7);341-345. https://doi.org/10.1016/S0165-6147(03)00159-7.

18 Kenett RS, Kenett DA. Quality by Design applications in biosimilar pharmaceutical products. Accreditation and Quality Assurance. 2008; 13 (12): 681–690. doi:10.1007/s00769-008-0459-6. S2CID 110606284.

19 https://www.juran.com/blog/quality-by-design-qbd-an-overview/ [Accessed January 2024]

20 Juran JM. Juran on Quality by Design: The New Steps for Planning Quality into Goods and Services. 1992 Free Press. Simon and Schuster

21 Yu LX. Pharmaceutical Quality by Design: Product and Process Development, Understanding, and Control. Pharmaceutical Research. 2008; 25 (4): 781–791. doi:10.1007/s11095-007-9511-1. PMID 18185986. S2CID 11700550.

22 Pharmaceutical Quality for the 21st Century: A Risk-Based Approach https://www.fda.gov/aboutfda/centersoffices/officeofmedicalproductsandtobacco/cder/ucm128080.htm [Accessed January 2024].

23 Rathore AS, Kapoor G. Implementation of Quality by Design for processing of food products and biotherapeutics, Food and Bioproducts Processing.2016; 99:231-243. https://doi.org/10.1016/j.fbp.2016.05.009.

24 Hassoun A, Jagtap S, Trollman H, Garcia-Garcia G, Alhaj Abdullah N, Goksen G et al., Food processing 4.0: Current and future developments spurred by the fourth industrial revolution, Food Control. 2023; 145: 109507. https://doi.org/10.1016/j.foodcont.2022.109507.

25 Teixeira JA, Vicente AA, Macieira da Silva FF, Azevedo Lima da Silva JS, da Costa Martins RM (2014) In: Teixeira JA, Vicente AA (eds) Engineering Aspects of Food Biotechnology, 1st edn. CRC Press, Boca Raton, USA

26 Hitzmann B, Hauselmann R, Niemoeller A, Daryoush Sangi D, Traenkle J, Glassey J Process analytical technologies in food industry ‐ challenges and benefits: a status report and recommendations. Biotechnol J. 2015; 10:1095-1100. https://doi.org/10.1002/biot.201400773.

27 Workman J, Lavine B, Chrisman R, Koch M. Process analytical chemistry Analytical Chemistry,2011; 81:4623-4643.

28 Varmuza K, Filzmoser P. Introduction to multivariate statistical analysis in chemometrics, CRC Press, 2009.

29 Chapman J, Power A, Netzel ME, Sultanbawa Y, Smyth HE, Truong VK et al. Challenges and opportunities of the fourth revolution: a brief insight into the future of food, Critical Reviews in Food Science and Nutrition, 2022; 62(10): 2845-2853. doi: 10.1080/10408398.2020.1863328

30 Djekic I, Mujčinović A, Nikolić A, Jambrak AR, Papademas P, Feyissa AH et al. Cross-European initial survey on the use of mathematical models in food industry. J Food Eng. 2019: 261:109-116. https://doi.org/10.1016/j.jfoodeng.2019.06.007

2ANALYTICAL METHODS

2.1 E-nose (HS-Ultra Fast GC-FID)

The e-nose technology simulates the human olfactory system. Typically, an electronic nose consists of an array of electronic chemical sensors (most often inorganic oxides) with partial specificity for some classes of volatile molecules. An appropriate pattern recognition system elaborates the overall signal and recognize the odour without any specific information on the perceived molecules [1]. In 2010 a breakthrough was made by Alpha MOS (Toulouse, France) that propose an e-nose based on the ultra-fast gas chromatography (UF-GC) technique [2]. This GC based e-nose (GC-FID e-nose) is spreading due to its use in a similar way to a classical e-nose, but with the possibility to obtain putative identification of the molecules present in the odour [3,4,5].

In particular, we used the instrumentation Heracles II ®, by Alpha MOS, Toulouse, France, implemented with an autosampler for headspace injection (PAL-RSI), a double-columns ultra-fast-chromatography system with two Flame Ionization Detectors (FID). The autosampler can condition the samples at controlled temperature before the head space collection to allow the concentration of the volatile molecules between sample and headspace to equilibrate. After injection, volatile molecules are collected in a Tenax trap and then released into the two chromatographic columns.

2.1.1 Basil aroma analysis

For basil analysis about 30 g of the whole basil plants, including leaves and stems, were exactly weighted at 0.1 g and hashed in a blender (Oster, Sunbeam Products Inc., Boca Raton, FL, USA) for 30 s in 300 mL of extraction solution at room temperature. The extraction solution was prepared with NaCl at a concentration of 100 g L^{-1}, to increase the volatiles release in the headspace (next step of the analysis), and 6 mg kg^{-1} of ethyl iso-butyrate to serve as internal standard for the CG analysis. After 30 s of resting time, 20 µL of the solution was collected and transferred in 20 mL amber vials that were immediately sealed and sent for analysis. Each extract was sampled at least three times in different vials. Samples vials were incubated for 20 minutes at 40°C, before injection with 500 rpm agitation (5 s on, 2 s off). Then 1 mL of air headspace was injected with a syringe temperature of 50°C.

2.1.2 Pesto aroma analysis

For the pesto analysis 2 grams were collected, transferred in a 20 mL vials and immediately crimped. Samples were then incubated at 50°C for 15 minutes with 500 rpm agitation (5 s on, 2 s off), then 5 mL of the headspace were collected and injected in the GC-FID e-nose.

In both cases, trap loading conditions were 18 s at 40°C, then flashed to 250°C for the release into the two columns at split ratio 1:1.

Columns have both length of 10 m, internal diameter 0.18 mm, film thickness 0.40 micron and are respectively MXT-5 (non-polar) and MXT-1701 (slightly polar).

For both analysis the temperature ramp for the two columns was 50°C for 2 s, then to 80°C at 1°C/s, then to 250°C at 3°C/s. The total time was 110 s. The carrier gas was hydrogen.

To calculate areas and concentrations of the volatile molecules AlphaSoft v16.0 software (Alpha MOS, Toulouse, France) was used.

For further chemometric data elaboration the raw chromatograms were exported in a suitable format for importing them in Matlab environment. In Figure 2-1 an example of an GC-FID e-nose chromatogram of pesto acquired with the XT-5 column is shown.

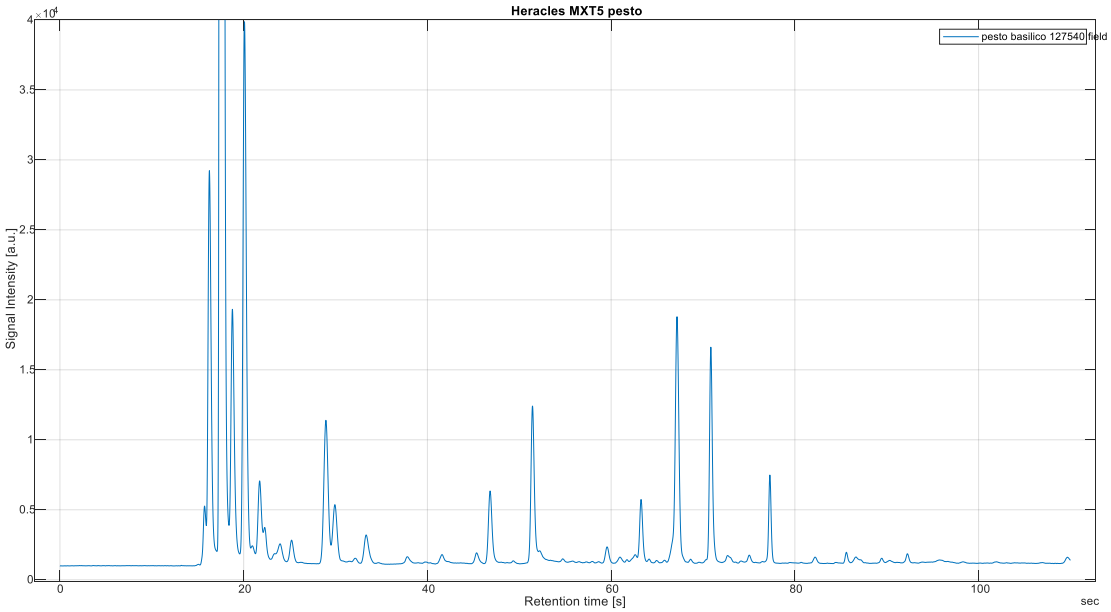


Figure 2-1. Chromatogram of a pesto sample on Heracles II MXT5 column.

2.2 Head-Space-Gas-Chromatography-Ion-Mobility-Spectrometry (HS-GC-IMS)

The IMS (Ion Mobility Spectrometry) technique [6] measures the time employed by a soft ionized molecule, accelerated by a uniform electric field, to reach the detector moving through an inert gas flow (nitrogen) at ambient pressure. This time depends on the ion mobility, that is characteristic of each molecule and depends on its mass and its steric hindrance. Molecules with different ion mobility can thus be separated and detected [7].

IMS instruments are extremely sensitive devices commonly used to detect drugs or explosive (i.e., at airport security checks). Due to the fast separation timescale (milliseconds) they are often used coupled to other techniques like mass spectrometry, gas chromatography or high-performance liquid chromatography to obtain a multi-dimensional separation [8].

In our case the analyses were performed by a FlavourSpec ® (G.A.S. mbH, Germany) GC-IMS instrument that use a GC column FS-SE-54-CB-0.5 (length 30 m, internal diameter 0.32 mm, film thickness 0.5 micron) for the first separation dimension. After the chromatographical separation the volatile molecules enter the drift tube where they are ionized reacting with reactant ions (water molecules naturally present in the drift tube charged by a β-radiation source of tritium). Ions are then accelerated towards the detector and the drift time is recorded.

Drift tube was maintained at ambient pressure and constant temperature of 80°C.

For the analysis of pesto 2 grams were transferred in a 20 ml glass vial and immediately crimped. The samples were equilibrated for 20 minutes at 60°C before the headspace collection by the autosampler. After that 1 mL of headspace was injected. The output is a landscape for each sample that reports on the x-axis the retention time, on the y-axis the drift time (related to the ion mobility) and on z-axis the signal intensity (Figure 2-2).

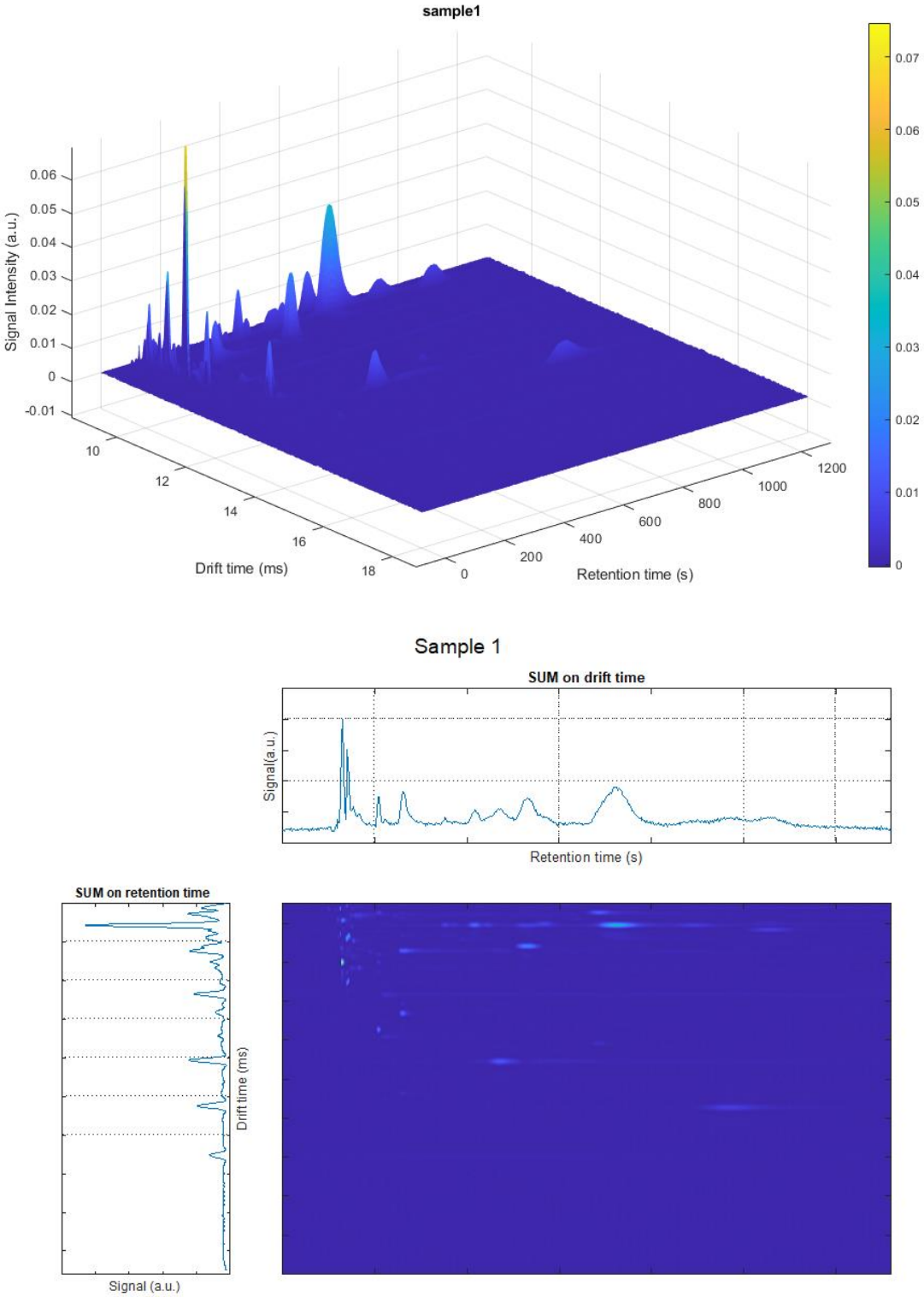


Figure 2-2. Upper part: 2D HS-GC-IMS chromatogram (pre-processed data). Lower part: peak map with the 2-dimension projected as sum on the relative axes, respectively on X axes the retention time, and on the Y axes the drift time.

2.3 Near InfraRed Spectroscopy (NIRS)

Spectroscopy study the absorption or emission of electromagnetic waves by matter. The measure of such radiations is a way to obtain information about the systems and its components and is called spectrometry. There are several spectroscopic techniques based on the different ways the electromagnetic radiation interacts with matter, depending on the energy of the radiation [9].

Literature reports numerous studies that present applications of spectroscopy in research as well as in industrial environment [10]. Low energy techniques are particularly useful in food analysis, because they are fast, non-destructive, not dangerous for the operators, easy to use and often do not require any sample preparation.

We will focus just on one of the techniques widely applied in food analysis that is the infrared spectroscopy, in mid or near infrared ranges. This technique uses the interaction of electromagnetic radiation with the vibrational states of covalent bonds and rotational states of molecules. For this reason, it is very powerful to measure foods that are composed by organic material containing covalent bonds between atoms like carbon, nitrogen, oxygen, sulphur, hydrogen.

Some characteristic regions in the NIR spectra which are linked to food components are reported in Table 2-1.

Table 2-1. Principal types of NIR absorption bands and their location in the spectrum

Wavelength interval	**Absorption bands**
800 – 1100 nm	N-H 2nd overtone OH 2nd overtone CH 3rd overtone
1100 – 1300 nm	CH 2nd overtone OH combination
1300 – 1420 nm	CH combinations
1420 – 1600 nm	OH 1st overtone NH 1st overtone
1600 – 1800 nm	CH 1st overtone
1800 – 2200 nm	OH combinations NH combinations
2200 – 2500 nm	CH combinations

In food analysis NIR spectroscopy is largely used to measure concentrations of some parameters after the calculation of proper calibration curves by multivariate techniques, for its rapidity and easiness of use. Moreover, the NIR instrumentation is becoming more and more small and cheaper opening to new possible applications [11,12] in situ and in-line. It is also possible to use the VIS-NIR spectra to have a qualitative description of the sample analysed in an untargeted approach, when the aim is to discriminate samples having some differences.

For the NIRS laboratory analysis undertaken for the work presented in this Thesis a benchtop DS-2500 (FOSS, Denmark) instrument was used.

Spectra acquisition of "*Pesto alla Genovese*" samples were done just transferring about 100 ml of pesto into the large cup of the instrument without any other preparation step.

Spectra were collected in the range from 400 to 2500 nm (8 replicates for each sample) and the raw spectra were exported for further statistical analysis.

In Figure 2-3 an example of spectra acquired on pesto samples is reported.

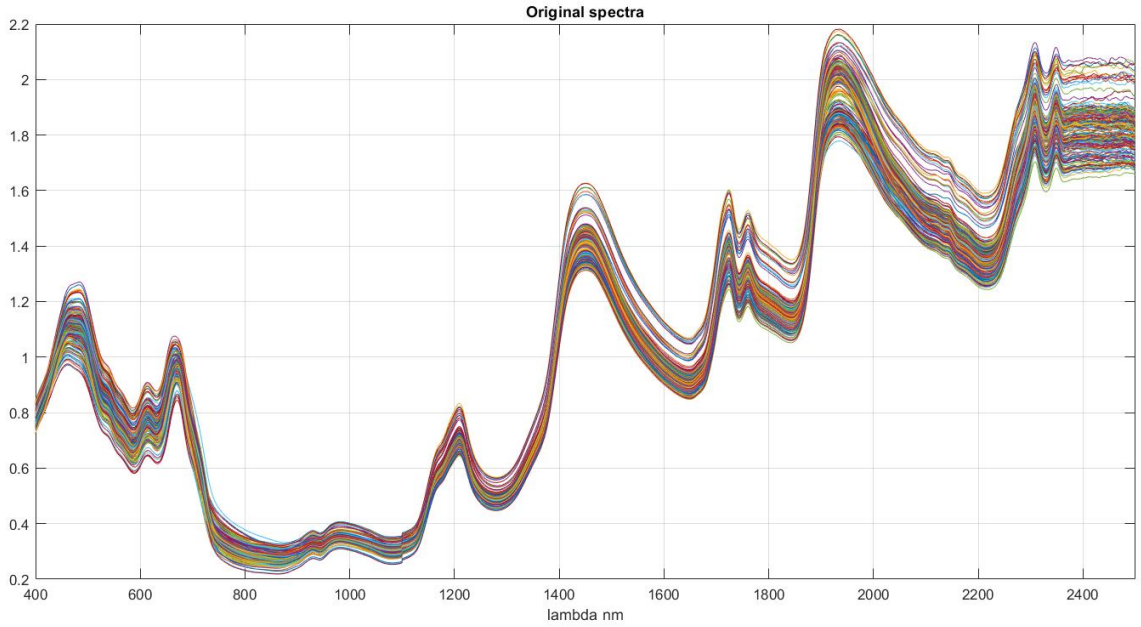


Figure 2-3. NIR raw spectra of pesto samples.

For the NIR analysis conducted on-line in the pesto production plant a Pro-Foss spectrometer (Foss, Hillerod, Denmark) was used with a spectral range from 1100 to 1650 nm with a resolution of 0.5 nm and 64 scans per sample. The instrument was equipped with an optical fibre that connect the spectrometer to the acquisition probe located on the process pipe.

2.4 Pesto stability analysis

"*Pesto alla Genovese*" is a complex multiphase system with an emulsion of oil in water, mixed with a watery cheese phase in which are suspended solid pieced like basil leaves and cashews pieces. Its equilibrium depends on the proper ingredient combinations and is stabilized by the emulsifying effect of the milk proteins. When this equilibrium is not stable the oil separation after some time is one of the effects that could be observed. Despite this oil release does not change the nature of the product it is not appreciated by consumers. Became so important to measure the physical stability of the "Pesto" system.

Stability has been evaluated by the LUMiSizer® (LUM, Berlin, Germany). It is basically a centrifuge equipped with a device for measuring the extinction of the transmitted light (NIR 856 nm and blue 470 nm) across the entire length of the cuvette sample in real time during the centrifugation process. It uses the STEP-Technology that permits to obtain Space- and Time-resolved Extinction Profiles over the entire cuvette holding the sample. Up to 12 different samples can be analysed simultaneously. Parallel light (L_0) illuminates the entire sample cell, and the transmitted light (L) is detected by two-thousands CCD sensors arranged linearly across whole sample cuvette from top to bottom, with a microscale resolution. Transmission is converted into extinction by taking the $\log(L/L_0)$ and the particle concentration can be estimated in each point of the cuvette. The speed of the centrifuge can be changed from 200 to 4000 rpm (corresponding in the middle of the cuvette to 5 to 2300 g). It allows to measure drops and particles velocity distribution for phenomena like creaming or sedimentation and so it is possible to have an estimation of a product stability and make shelf-life prediction.

The instrument can control the temperature from 4 to 60 °C.

In the present work the method used for the pesto characterisation used temperature 30°C, rotor speed 4000 rpm, light 865 nm, cuvette PA 10 mm optical path.

2.5 Spectral Imaging

2.5.1 Visible Red Green Blue (Vis-RGB) imaging

The digital image processing was born in the 1960s on satellite images, mainly with the contribution of the Bell Laboratories, Massachusetts Institute of Technology, and the Maryland University [13].

Initially image processing consisted of methods dedicated to improving the image quality, in fact the first digital images had very poor quality. The Jet Propulsion Laboratory (JPL) used image processing tools to improve image quality and to extract information from the images sent back by the Space Detector Ranger 7 in 1964. From then on, the increased quality of photographic sensor joined to the elaboration power has started a new discipline that extend its application to many fields, from medicine to food analysis [14,15,16]. An idea of how much these applications are spreading is given by the number of reviews published in 2023, just on “image processing and food”, that overcome the 600.

In food analysis, despite its simplicity, colour analysis plays a big role, since colour change can be the results of oxidation and decomposition processes thus capturing, albeit indirectly the “chemistry“ of food. Moreover, texture and appearance are important sensory attributes.

In this work a Red Green Blue (RGB) camera has been used for basil characterisation, and a hyperspectral camera for pesto characterisation as will be shown in paragraph 6.1 and 6.2 respectively.

A vision system produced by SENSURE (SENSURE SRL, Bergamo, Italy) was used to acquire RGB images (24-bit, resolution 1280x1020 pixels). In Figure 2-4 it is shown as example of one of the basil images acquired in-line.

The vision system software automatically extracts few features from the images and store them.

Figure 2-4 Example of a basil image acquired by the on-line RGB camera.

2.5.2 Hyper Spectral Imaging (HSI)

Whereas the human eye sees colour in the visible part of electromagnetic spectrum, mainly in three bands (around red, green, and blue), hyperspectral imaging collects for each pixel of an image a large electromagnetic spectrum with fine wavelength resolution, covering often a spectral range from ultraviolet to near infrared. In this way, it is possible to obtain much more information.

Hyperspectral imaging [17,18] (HSI) was first applied in the mining and geology field for its ability to identify minerals or soils characteristics, but rapidly HSI applications spread to many other fields, mainly with the development of instruments installed on board of artificial satellites. Some of the fields range from agriculture to ambient protection, to biomedical to astronomy. In recent years also to food analysis, processing and controlling.

Its powerfulness derives from the possibility to give simultaneously morphological and chemical information.

In my period spent at the INRAn facilities in Montpellier (France) [19] pesto images acquisition was done using two separate hyperspectral cameras (see Figure 2-5):

- Vis-NIR HSNR03 camera (wavelength 409 - 987 nm)
- NIR HSNR05 camera (wavelength 964 - 2494 nm).

Both cameras acquire the image in line scan (one row at time) modality.

Pesto samples were acquired in small aluminium cup sampling from the middle of the jar, including in each image as reference a white plate (see Figure 2-6).

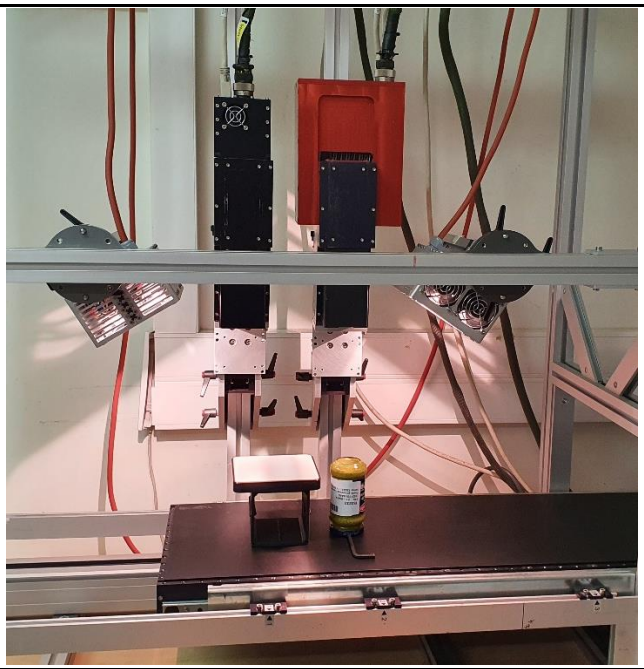


Figure 2-5. Hyperspectral cameras at INRAn facilities.

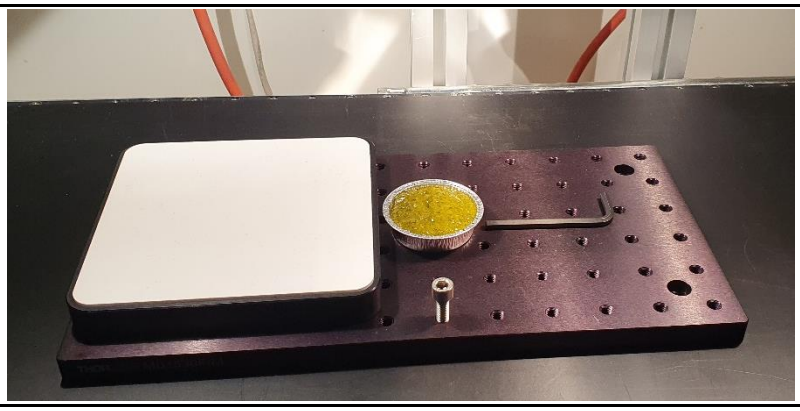


Figure 2-6. Pesto sample acquisition with reference white plate.

1 Wei G, Dan M, Zhao G, Wang D. Recent advances in chromatography-mass spectrometry and electronic nose technology in food flavor analysis and detection, Food Chemistry, 2023; 405 Part A, 134814. https://doi.org/10.1016/j.foodchem.2022.134814.

2 https://growfers.com/story/alphamos/ [Accessed January 2024]

3 Wasilewski T, Migoń D, Gębicki J, Kamysz W, Critical review of electronic nose and tongue instruments prospects in pharmaceutical analysis, Analytica Chimica Acta.2019;1077: 14-29. https://doi.org/10.1016/j.aca.2019.05.024

4 Śliwińska M, Wiśniewska P, Dymerski T, Wardencki W, Namieśnik J., 8 - Advances in Electronic Noses and Tongues for Food Authenticity Testing, Editor(s): Gerard Downey, In Woodhead Publishing Series in Food Science, Technology and Nutrition, Advances in Food Authenticity Testing, Woodhead Publishing, 2016, Pages 201-225, https://doi.org/10.1016/B978-0-08-100220-9.00008-4.

5 Damiani T, Cavanna D, Serani A, Dall'Asta C, Suman M., GC-IMS and FGC-Enose fingerprint as screening tools for revealing extra virgin olive oil blending with soft-refined olive oils: A feasibility study, Microchemical Journal. 2020; 159: 105374, https://doi.org/10.1016/j.microc.2020.105374.

6 Parastar H, Weller P, Towards greener volatilomics: Is GC-IMS the new Swiss army knife of gas phase analysis?, TrAC Trends in Analytical Chemistry. 2024; 170: 117438 https://doi.org/10.1016/j.trac.2023.117438.

7 Gu S, Zhang J, Wang J, Wang X, Du D, Recent development of HS-GC-IMS technology in rapid and non-destructive detection of quality and contamination in agri-food products, TrAC Trends in Analytical Chemistry. 2021; 144: 116435 https://doi.org/10.1016/j.trac.2021.116435.

8 Baglai A, Gargano AFG, Jordens J, Mengerink Y, Honing M, van der Wal S et al. Comprehensive lipidomic analysis of human plasma using multidimensional liquid- and gas-phase separations: Two-dimensional liquid chromatography–mass spectrometry vs. liquid chromatography–trapped-ion-mobility–mass spectrometry, Journal of Chromatography A. 2017; 1530: 90-103 https://doi.org/10.1016/j.chroma.2017.11.014.

9 Osborne, BG, Fearn T, Hindle PH Practical NIR spectroscopy with applications in food and beverage analysis. Longman Scientific and Technical, Harlow, 1993

10 Scotter CNG. Non-destructive spectroscopic techniques for the measurement of food quality, Trends in Food Science & Technology. 1997; 8: 285-292 https://doi.org/10.1016/S0924-2244(97)01053-4.

11 Rodriguez-Saona L, Aykas DP, Rodrigues Borba K, Urtubia A, Miniaturization of optical sensors and their potential for high-throughput screening of foods, Current Opinion in Food Science. 2020; 31: 136-150 https://doi.org/10.1016/j.cofs.2020.04.008.

12 Gopal J, Muthu M, Handheld portable analytics for food fraud detection, the evolution of next-generation smartphone-based food sensors: The journey, the milestones, the challenges debarring the destination, TrAC Trends in Analytical Chemistry. 2024; 171:117504 https://doi.org/10.1016/j.trac.2023.117504.

13 Rosenfeld A. ACM Computing Surveys. 1969; 1: 147–176 https://doi.org/10.1145/356551.356554

14 Meenu M, Kurade C, Neelapu BC, Kalra S, Ramaswamy HS, Yu Y, A concise review on food quality assessment using digital image processing, Trends in Food Science & Technology. 2021; 118: 106-124 https://doi.org/10.1016/j.tifs.2021.09.014.

15 Jackman P, Sun DW. Recent advances in image processing using image texture features for food quality assessment. Trends in Food Science & Technology.2013; 29(1): 35-43 https://doi.org/10.1016/j.tifs.2012.08.008.

16 Mahanti NK, Pandiselvam R, Kothakota A, Ishwarya P SP, Chakraborty SK, Kumar M, et al. Emerging non-destructive imaging techniques for fruit damage detection: Image processing and analysis, Trends in Food Science & Technology. 2022; 120: 418-438 https://doi.org/10.1016/j.tifs.2021.12.021.

17 Amigo J, Babamoradi H, Elcoroaristizabal S. Hyperspectral image analysis. A tutorial Analitica Chimica Acta. 2015; 896: 34-51

18 Hyperspectral Imaging. in Data Handling in Science and Technology series Ed. J. Amigo, 2020 Elsevier B.V. https://doi.org/10.1016/B978-0-444-63977-6.00001-8

19 https://www.chemproject.org/chemhouse [Accessed January 2024]

3CHEMOMETRICS METHODS

3.1 Few words about Chemometrics

Chemometrics is an interdisciplinary science that combines statistics and chemistry. It is practically oriented to solve analytical chemistry problems (and not only) using advanced statistical tools [1,2,3,4].

Called an “art” [5] by Svante Wold, one of the founders of this discipline, chemometrics helps to extract relevant information from chemical data. In fact, in analytical chemistry is quite common to have, as results of experiments, a large quantity of data in which noise and useful information are mixed.

Born in the early 1970s and facilitated on one hand by the increase in the computer power and, on the other hand, by the analytical instruments’ development its use has been largely spread for more than one reason. The most relevant is the augmented consciousness that chemometrics is not a "facultative” appendix, but a fundamental everyday tool [6,7,8,9,10].

Jus to mention an example, chemometrics give a relevant contribution on the design of the experimental trials, where it overcomes the old (but still strongly rooted) idea of changing "one variable at time”. Nature is a multivariate system and so it is crucial to have proper tools able to manage this complexity. Chemometrics does that using a multivariate approach to data analysis.

Possibilities are a lot and some of them have been explored in this Thesis.

Examples of classification techniques will be reported, useful when like in our case, a comparison between some classes of samples is pursued. In our cases, we had some additional information on the systems we were studying (i.e. the different recipes of a food product) and typically “*supervised*” models were applied. In other cases, we did not have additional information and so “*unsupervised*” methods will be required.

3.2 Data pre-processing

Data pre-processing [11,12,13,14] is a fundamental step needed to remove noise or sources of variability which are not inherent to the sought information, e.g. related to the physical characteristic of samples when compositional profile is of interest, variability due to ambient conditions, variation in instrumental settings, etc. In general, it can be distinguished signal pre-processing (applied in the row direction sample by sample) from pre-processing such as centring and scaling (applied in the columns directions of the dataset) [15]. Here, are concisely reported the signal pre-processing applied per type of signal.

Imaging pre-processing is described in the paragraph 3.6.

3.2.1 Chromatographic data

Chromatographic data may be affected by retention time shift from run to run and when the chromatograms are analysed as such, i.e. without peak recognition and integration, this represents an issue for further multivariate data analysis, as well it does baseline disturbance. In addition, normalization may be needed to compensate run to run intensity variability, and the presence of major and minor components may require scaling to let all of them to contribute fairly to the modelling phases.

In the GC-FID e-nose analysis an Internal Standard (IS) was used in each chromatographic run and thus normalization was applied by dividing data by the IS signal.

Also, the gas chromatograms acquired by the Heracles II instrument, despite their high reproducibility and stability, showed both baseline and retention times shift.

Thus, retention time samples alignment was done by using the Interval Correlation Optimised Shifting algorithm (icoshift) applied by intervals, which were manually defined.

The icoshift algorithm was initially proposed in 2010 by Tomasi et al. [16] for NMR spectral data, and then extended in 2011 to chromatographic data [17]. It is based on COrrelation SHIFTing of spectral intervals and employs an FFT engine that aligns all spectra simultaneously. The algorithm is demonstrated to be faster than similar methods found in the literature making full-resolution alignment of large datasets feasible.

Baseline subtraction was operated by using the weighted least squares algorithm (2nd order polynomial) [18, pages 173-174].

Finally, since the peaks' intensity and variance reflect the presence of major and minor constituents, it was important to use a procedure able to make the different chromatographic regions influence on the developed statistical models comparable. To this aim block scaling to equal block variance (defining the blocks to be the same as the intervals used for the alignment with icoshift) was used, including column mean centring.

3.2.2 Spectroscopic data

The name spectroscopy encompasses many techniques depending on the wavelength, and so the energy, used. In fact, the energy of the light that interacts with the matter causes phenomena related to absorption of energy at atomic level causing changes in electronic state (X-ray and UltraViolet Visible absorption), or molecular level, with changes in rotational and vibrational states (InfraRed and Raman) or in rotational states (microwaves and Nuclear Magnetic Resonance).

We will focus just on Near Infrared Spectroscopy (NIRS). This technique, despite the infrared radiation was discovered in 1800, show its first practical applications starting just in late 1960s, mainly for moisture determination [19,20,21,22]. This late spread could be attributable to the lack of instruments, but also to the absence of proper mathematical tools to extract analytical information from the spectra. In fact, in the NIR range the Lambert-Beer law (that relates linearly the light absorption of an absorbing analyte with its concentration) is not applicable.

Moreover, in the NIR spectroscopy there are several 'disturbing' factors, like light scattering phenomena, overlapped signals, background effects, bands overtones (with internal correlations of signals) and the absorption of water (almost ubiquitarian in food systems) in a wide part of the spectrum.

All these considerations explain why is necessary to pre-process NIRS data before extracting relevant information.

Spectral (or signal more in general) preprocessing is itself a field of research, and detailing it is beyond my aim. In general, in NIRS preprocessing may be divided into three main categories smoothing, baseline correction, and normalization [13,23]

The preprocessing applied in my Thesis work is reported below:

- Smoothing by Savitzky-Golay filter (SG). This filter removes the high frequency noise by polynomial interpolation (codified in specific filter) applied by spectral window (a zero-degree polynomial corresponds to moving average)
- Transforming the signal to its first or second Derivate, applied on a smoothed signal so to remove noise. First derivative can remove constant background. Second derivative removes constant and additive background. In addition, implicitly they can deconvolute to some extent overlapped band by highlighting the presence of shoulders, etc.
- Normalization by the SNV (Standard Normal Variate) method. This is done to make all spectra comparable, passing to relative intensities (or absorbance level). It can be useful to correct spectra for changes in optical path length and light scattering (it is assumed that the standard deviation of the spectra represents well these changes). SNV is, for example, frequently used to compensate for changes in surface roughness of the material [23]. Mathematically, SNV consists in subtracting each spectrum by its own mean and dividing it by its own standard deviation, so after SNV each spectrum will have a mean of zero and a standard deviation of one.

- Multiplicative Scatter Correction (MSC). In this case, it is assumed that chemical variation is small compared to physical variation (i.e. variation introducing a constant (additive)/proportional (multiplicative) baseline effect) and thus the true 'signal' may be replaced by a constant reference signal, usually the mean (or median) spectrum, m (it may also be a specific spectrum of the data set).

3.3 Exploratory data analysis and modelling methods

3.3.1 Principal Component Analysis (PCA)

Principal Component Analysis [24,25,26] is an important and powerful method used for explorative analysis of dataset containing high number of dimensions for each observation.

It increases the interpretability and visualization of multidimensional data while preserving the maximum amount of information.

Originally proposed by Pearson in 1901 and subsequently improved by Hotelling in 1933, it became computationally feasible to use on larger dataset after the availability of computers.

Fundamentally PCA reduce the dimensionality of a dataset linearly transforming the data into a new orthogonal coordinate systems (Principal Components) where most of variation in the data can be described.

PCA works finding a new reference space (hyperspace) in which the centre is the average value of the original data; then the first principal component direction is calculated from the centre in the direction that maximize the data variance. The second component is calculated in the direction, orthogonal to the first, that again maximize the data (residual) variance. The process is repeated for each principal component.

Mathematically PCA is represented as (Figure 3-1)

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{T} + \boldsymbol{E} \qquad \textit{(Eq. 3-1)}$$

where:
X is the original dataset
T is the score matrix
P is the loadings matrix
E is the matrix of the residuals.

The scores matrix **T** describes how the different rows in **X** (observations) relate to each other. Scores are the coordinates of samples in the PCA space (i.e. each scores vector is a linear combination of original variables). The scores plot is a powerful tool to display patterns in multivariate data.

The loading matrix **P** holds the weights of the linear combination and thus reflect the influence of the variables in **X** in defining the PCA model. In other words, loadings indicate which variables are responsible for the pattern found in scores **T**.

The loadings plot shows graphically how the variables are related.
Discussing the scores and loadings plots jointly allows linking pattern observed in scores plot to the variables responsible for them.

The residual matrix **E** is the noise part of the data. It represents the part of **X** not explained by the model $\mathbf{TP}^{\mathbf{T}}$. Plotting the changes in residual variance vs the number of PCs is one of the criteria which can be used to establish the best number of PCs (scree plot, introduced by Cattel in 1966 [27]).

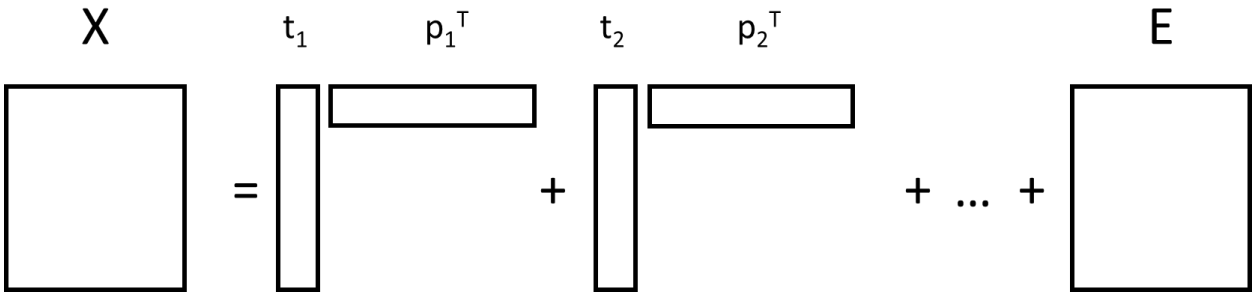


Figure 3-1. Principal Component Analysis scheme: **t**1 and **p**1 are respectively the scores and the loadings of the first PC, and so on. **E** is the residual matrix.

3.3.2 Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS)

In the cases in which the measured data are the results of a combination of different contributions and the interest is recovering them distinctly, then resolution/spectral unmixing methods can be applied. For examples in spectroscopy where a spectrum is the combination of the spectra of the pure components of a mixture, or in chromatography where the signal intensity of a chromatogram is the combination of the signals (partially overlapped) of singles molecules, or in hyperspectral imaging where to each pixel correspond a spectrum resulting from the combination of the pure spectra of the individual components present in the system.

Among the different methods, in this Thesis works we applied Multivariate Curve Resolution Alternating Least Squares [28,29,30], which is a curve resolution method assuming the data follow a bilinear model, that is the observed signal (spectrum or chromatogram or other) is a linear combination of the pure components in the system.

MCR-ALS decomposes the **D** data matrix into the product of matrices **C** (the concentrations of each resolved component in the samples) and $\mathbf{S}^T$ (the spectra profile of each resolved "pure" component).

The bilinear model could be written as (Figure 3-2):

$$D = CS^T + E \qquad \textit{Eq. 3-2}$$

where:
D is the original dataset
C holds the relative concentrations of the "pure" components
$\mathbf{S}^T$ holds the spectral profile of each "pure" component.
E is the unmodeled part of the data **D**

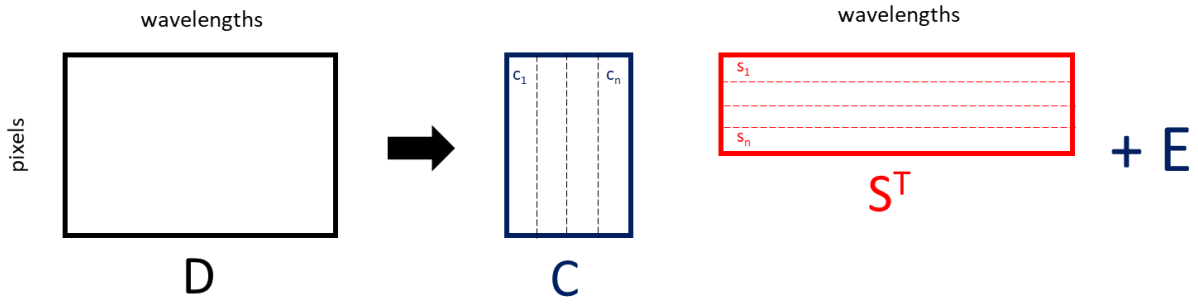


Figure 3-2. MCR-ALS example of application for a hyperspectral image. In matrix **C**, $\mathbf{c_1}$ to $\mathbf{c_n}$ are the components in which **D** has been decomposed and $\mathbf{s_1}$ to $\mathbf{s_n}$ are the correspondent pure spectra.

To apply MCR some constrains need to be imposed to give unique results, in fact the ALS solution suffers from rotational ambiguity. Constrains to give proper results should be consistent with the nature/behaviour of the studied systems. Typically for spectral systems non-negativity constrains are applied on both concentration and spectra dimensions, within the assumption that a concentration as well as spectral signal should not be negative. Other possible constraints that could be applied are unimodality, that forces a "pure" component to be constituted of just one peak (it can apply to chromatographic signals), or closure (i.e. the concentration of the resolved components in each sample should close to 1 or 100 %) in the case of mixture data. Other useful constraints, which can be applied in a flexible way to just one or all the components, are selectivity constraints, which use a priori knowledge on the spectral profile of pure species, e.g. imposing zero in region where they do not absorb.

The number of components to be used in the decomposition should be chosen carefully. Typically, a knowledge of the chemistry of the system could help, otherwise several MCR models with different numbers of components could be tried and for each of them the interpretability of its resolved spectral profile, based on the pure spectrum to which could be associated, or the presence of bands that are meaningful with respect to the composition of the studied system, should be evaluated. A suggested rule of thumb is that if two models provide equally plausible solutions the solution with least components will be chosen [30].

3.3.3 ANOVA-simultaneous component analysis (ASCA)

Analysis of variance (ANOVA) is a method applied to designed data (i.e. data acquired by systematic varying one or more conditions at specified levels) to assess the effect of the experimental factors, e.g. different samples categories, treatments, etc., on each dependent variable. However, ANOVA does not suffice to analyse multivariate data since it does not take the interrelation between variables into account. The classical extension of ANOVA to multivariate data is multivariate-ANOVA (MANOVA) [31]. However, MANOVA is not able to analyse data when the number of variables exceed the number of measured samples (example in the case of a spectra) also, multinormal distribution of the data is assumed, which is rarely fulfilled in complex dataset.

One of the methods proposed to overcome this limitation is ANOVA-simultaneous component analysis (ASCA) [32,33]. In the ASCA methodology ANOVA is merged with PCA, removing in this way the drawbacks of both methods.

The formulation, in case of two studied factors, e.g. in agronomic studies plant variety and harvesting season, is as in equations 3.3 and 3.4. At first step, as in ANOVA, the data matrix $\boldsymbol{X}$ is partitioned into the contribution of each factor and their interactions:

$$\boldsymbol{X_c} = \boldsymbol{X} - 1\boldsymbol{m}^T = \boldsymbol{X_1} + \boldsymbol{X_2} + \boldsymbol{X_{1x2}} + \boldsymbol{X_{res}} \qquad \textit{Eq. 3-3}$$

where $\boldsymbol{X_c}$ is the centred data matrix, $\boldsymbol{m}^T$ is the vector of column averages, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are the main effect matrices holding the levels average for factor 1 and 2 respectively, $\boldsymbol{X_{1x2}}$ is the interaction effect matrix and $\boldsymbol{X_{res}}$ is the residuals matrix.

Then, at a second step a Simultaneous Component Analysis (SCA) is performed, obtaining a scores matrix $\boldsymbol{T}$ and a loadings matrix $\boldsymbol{P}$ for each effect and interaction matrix:

$$\boldsymbol{X_c} = \boldsymbol{T_1P_1} + \boldsymbol{T_2P_2} + \boldsymbol{T_{1x2}P_{1x2}} + \boldsymbol{X_{res}} \qquad \textit{Eq. 3-4}$$

where $\boldsymbol{T}$ holds the scores and $\boldsymbol{P}$ the loadings of each PCA model; the maximum number of PCs for each model is equal to the number of levels minus one. In Figure 3-3a schematic representation of ASCA is shown

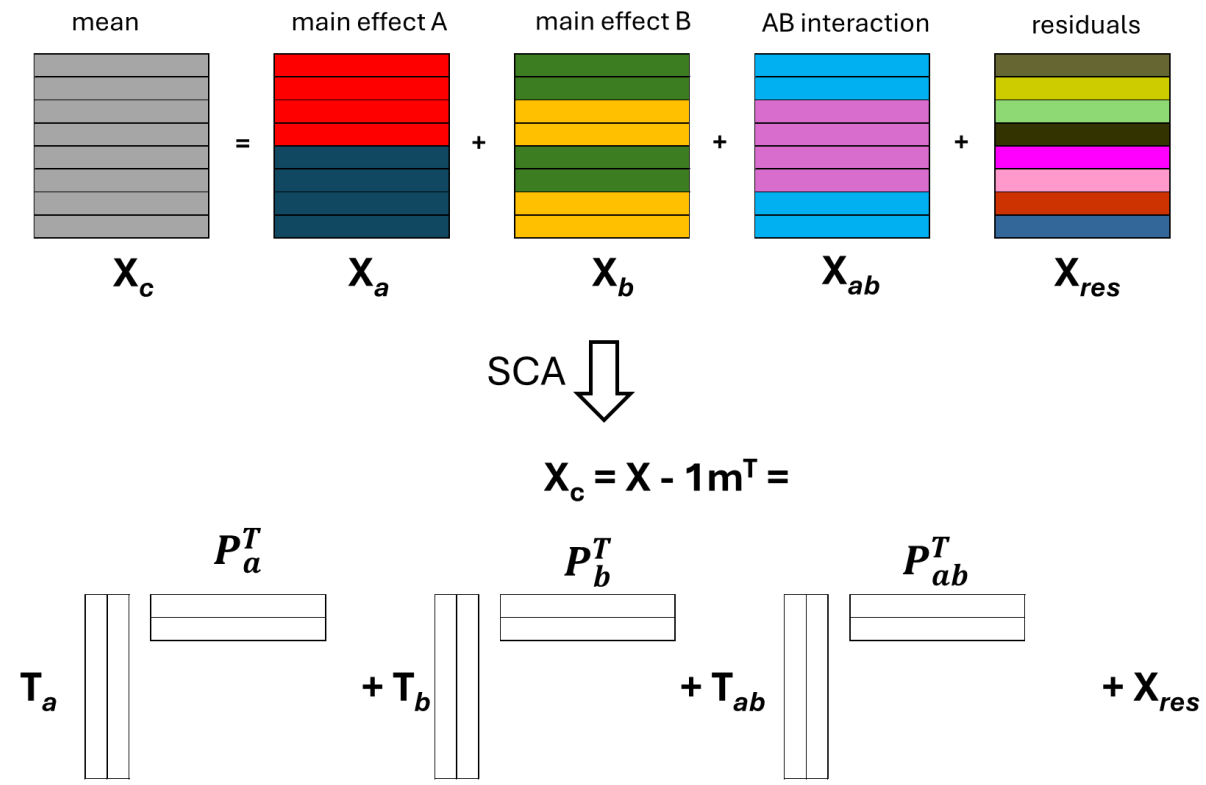


Figure 3-3. Scheme of ASCA.

To better inspect the ASCA results, *i.e.*, to highlight how the samples are dispersed around the mean of each effect level, could be useful to project the single sample on the ASCA scores plot. This can be achieved by adding the residuals to the estimated x_i values and then obtaining the single sample scores form the SCA model. For example, for each factor or interaction *(f)* the computation of the score vector $t_{i+res}(f)$ is carried out through the following equation:

$$t_{i+res}(f) = (X_i(f) + X_{res})p_{res}(f) \qquad Eq.\ 3\text{-}5$$

where $X_i(f)$ is the effect matrix for a specific factor or interaction and X_{res} is the residuals matrix, whereas $p_{res}(f)$ represents the loadings vector of the SCA model for the effect of that factor or interaction.

3.3.4 Partial Least Squares regression (PLSR)

Partial Least Squares Regression (PLSR) [34] is a widely used method for calibration and regression tasks.

The aim of PLS is to relate two sets of data, ***X*** and ***Y***, building a multivariate model based on maximization of ***XY*** covariance, and then use this model for prediction.

PLS overcome the limitations (collinearity issue and requirement of number of samples larger than number of variables) of multilinear regression by a first step of data compression by latent variables.

Two main algorithms can be used for the calculation of the PLS models, the NIPALS [35] and the SIMPLS [36] algorithms.

NIPALS, developed by H. Wold, calculates scores **T** and loadings **P** for the **X** block, and scores **U** and loadings **Q** for the **Y** block in a PCA-like way:

$$X = TP^T + E \qquad Eq.\ 3\text{-}6$$

$$Y = U Q^T + F \qquad \text{Eq. 3-7}$$

Where ***T*** and ***P*** are the scores and loadings, respectively for the ***X*** block while ***U*** and ***Q*** are the score and loadings for the ***Y*** block. ***E*** and ***F*** are the residual for ***X*** and ***Y*** blocks, respectively.

Specific to PLS is a weight matrix ***W*** that ensures maximization of the covariance between ***T*** and ***U***. In NIPALS algorithm the following calculation is repeated iteratively until convergence (component wise):

$$T = X W^T \qquad \text{Eq. 3-8}$$

$$W = U^T X \qquad \text{Eq. 3-9}$$

After the significant PLS components are calculated, by post processing the "pseudo-regression" coefficients matrix **B**, which relates the predictors **X** with the responses **Y** is calculated as:

$$B = W(P^T W)^{-1} Q^T \qquad \text{Eq. 3-10}$$

In this way, the model can be, finally, re-expressed in term of the original variables (which is useful for prediction):

$$\hat{Y} = B X \qquad \text{Eq. 3-12}$$

The second algorithm, SIMPLS, differs from the first one mainly in the way the ***X*** matrix is deflated after the first component. In this case a non-iterative approach uses Singular Value Decomposition of the covariance matrix $X^T Y$ to calculate loadings and scores.

3.4 Discriminant analysis

The possibility to understand if a certain sample is a part of one or more known categories falls under the general topic of classification. In terms of data analysis, it means that some mathematical/statistical rules will be defined as to assign each sample to one or more categories, based on the variables that describe the sample.

In this case, differently to exploratory analysis, is necessary to know a priori information about the categories, and this information is used to build the model (supervised method).

There are many multivariate classification methodologies, and to go into details for all of them is out of the scope of this Thesis. It is important to underline that they are divided into two main groups: aimed at discrimination and aimed at class modelling. In the first case the classification rules are defined to finds differences between sample categories (classes). This means to find directions in the geometrical space (hypersurface) of variables which allow assigning the samples of a given class to a specific region of the variables space. So, there will be defined as many regions as the number of classes. In class modelling instead the objective is on modelling similarity between samples that belong to the same class, and not to differentiate the classes. In this case the class modelling rules define the space of the class (hypervolume) without considering if there are other classes, and in an independent way from each other.

To resume, the main difference between discriminant classification and class-modelling, is that in the first case a sample is always assigned to one of the defined categories, while in the second case a sample could be assigned to one, more than one, or none of them.

In the case of discriminant classification, the decision rules to assign a sample to a given category are based on probability criteria (Baye's rule says that a sample is assigned to the class where it has the maximum probability to belong). However, under this general framework are comprised several methods. Among them, the most used are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Partial Least Squares Discriminant Analysis (PLS-DA). The first two share the limitations that they can be applied only when the number of samples is larger than the number of variables, and when variables have low correlation with each other. Both issues could be overcome by applying before discrimination a data reduction methodology providing an orthogonal subspace, like PCA and PLS-DA.

Hereafters are briefly recalled the discriminant techniques used in this thesis work, which are based on extension of PLS regression to discriminant classification task.

3.4.1 Partial Least Square – Discriminant Analysis (PLS-DA)

PLS-DA is a variant of partial least squares regression (PLS-R) that is used when the matrix of response variables ***Y*** is categorical (with discrete values) [37,38,39,40].

In Partial Least Squares Discriminant Analysis (PLS-DA) the discriminant classification problem is reformulated as a regression problem in which the responses matrix ***Y*** contains the class membership information in a binary coded form, *i.e.* each Y-column refers to a given category and each sample will have a value of 1 in the Y-column corresponding to its true class and zeros, or -1, elsewhere. Accordingly, the classification problem can be reformulated as finding the best regression model (by using the PLS algorithm) linking the experimental data measured on the samples (***X***) to the binary-coded dummy matrix ***Y***.

It has been demonstrated that PLS-DA converge to linear discriminant analysis if the number of PLS latent variables is equal to the number of variables in ***X***. The regression coefficients matrix (***B***, see eq. 3-13 above) allows prediction of the ***Y*** values for unknown samples $\boldsymbol{X}_{\text{new}}$; as the predicted values ($\hat{\boldsymbol{Y}}_{new}$) can assume real values, and not only ones and zeros, in this case a classification rule to assign the samples to a given category must be defined. In general, there are two approaches a "true" discriminant one where classification is accomplished by assigning the samples to the category corresponding to the highest value of the predicted dummy response, e.g. if the classification problem regards three classes, a sample whose predicted Y values are [0.98 0.5 0.1] will be assigned to class one. This approach when modelling more than two categories may be sub-optimal and it is suggested to apply LDA (or QDA) on the ***Y*** scores or on the ***Y*** predicted values [39], instead.

A second approach is based on the choice of a class threshold for each category [18], *i.e.* a value for each dummy **y**, if the predicted y for a sample is above it, then the sample is assigned to the class and *viceversa* if it is under. The threshold is usually chosen based on classification performance estimated in cross-validation.

3.4.2 Sequential and Orthogonalized - Principal Least Square (SO-PLS) and Sequential and Orthogonalized - Principal Least Square – Linear Discriminant Analysis (SO-PLS-DA)

SO-PLS is a multiblock extension of the PLS regression [41,42] in which the information is extracted sequentially from each predictor block and where the subsequent blocks are orthogonalized to the previously selected components. Unlike multiblock PLS where block scaling is essential because blocks are used altogether, block scaling is of no concern in SO-PLS. However, the order in which the blocks are presented to the algorithm can influence the results. The significance of the addition of any predictor block can be tested.
Considering, e.g. two blocks of predictors the SO-PLS steps are:

a. starts by one of the blocks, e.g. ***X***, and fit a standard PLS model. Thus obtaining

X-scores *(**T**$_X$),* the ***X***-weights (***W***$_X$), the ***X*** and ***Y*** loadings *(**P**$_X$ and **Q**$_X$* respectively) and ***Y***-residuals (***E*** = ***Y*** − ***T***$_X$***Q***$_X^T$)

b. The second block ***Z*** is orthogonalized with respect to the scores of the previous PLS model:

$$\boldsymbol{Z}_{orth} = \boldsymbol{Z} - \boldsymbol{T}_X (\boldsymbol{T}_X^T \boldsymbol{T}_X)^{-1} \boldsymbol{T}_X^T \boldsymbol{Z} \qquad \textit{Eq. 3-11}$$

c. ***Z***$_{orth}$ is then selected to calculate a PLS model with the ***Y***-residuals (***E***). Thus, obtaining the ***Z***$_{orth}$-scores (***T***$_{Zorth}$), the ***Z***$_{orth}$ -loadings (***P***$_{Zorth}$), the ***Z***$_{orth}$-weights (***W***$_{Zorth}$), and the ***Y***-loadings (***Q***$_{Zorth}$).

In this way, further information is extracted from ***Z*** that explains the remaining variance in ***Y,*** but which is orthogonal to the information previously contributed by block **X** (i.e. SO-PLS focus on the distinctive information each block carries).

d. In the last step the full predictive model is obtained adding the two models:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{T}_X \boldsymbol{Q}_X^T + \boldsymbol{T}_{\boldsymbol{Z}_{orth}} \boldsymbol{Q}_{\boldsymbol{Z}_{orth}}^T \qquad \textit{Eq. 3-12}$$

As in any PLS model this can be rearranged to be expressed in terms of regression coefficients:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{B}_{\boldsymbol{X}} + \boldsymbol{Z}_{orth}\, \boldsymbol{B}_{Zorth} \qquad \textit{Eq. 3-13}$$

Where:

$$\boldsymbol{B}_{\boldsymbol{x}} = \boldsymbol{W}_{\boldsymbol{x}} (\boldsymbol{P}_{\boldsymbol{x}}^{\boldsymbol{T}} \boldsymbol{W}_{\boldsymbol{x}})^{-1} \boldsymbol{Q}_{\boldsymbol{x}}^{\boldsymbol{T}} \qquad \textit{Eq. 3-14}$$

$$\boldsymbol{B}_{\boldsymbol{Z}_{orth}} = \boldsymbol{W}_{\boldsymbol{Z}_{orth}} (\boldsymbol{P}_{\boldsymbol{Z}_{orth}}^{\boldsymbol{T}} \boldsymbol{W}_{\boldsymbol{Z}_{orth}})^{-1} \boldsymbol{Q}_{\boldsymbol{Z}_{orth}}^{\boldsymbol{T}} \qquad \textit{Eq. 3-15}$$

The number of latent variables is decided independently for each block, usually by cross-validation.

Adding more blocks than two can easily be done by repeating orthogonalization with respect to the scores of the previous PLS regression and fitting the orthogonalized block with the preceding residual matrices.

Extending SO-PLS to the discriminant case (SO-PLS-DA or SO-PLS-LDA depending on the classification rule adopted [41]) can be done by using a *dummy* ***Y*** matrix containing the information about the class membership, as when passing from PLS to PLS-DA.

The CovSel features selection method [see section 3.5.1] has been also implemented in the SO-PLS framework, considering the multiblock nature of the method. Hereafter are reported the algorithm steps in the case of two predictor blocks ***X*** and ***Z*** and a dummy vector ***y*** (codifying two classes):

1. Variables are selected by Cov-Sel from **X** (as in standard Cov-Sel) and stored in matrix ***X***$_{sel}$
2. ***y*** is fitted to ***X***$_{sel}$ by Ordinary Least Square (OLS): $\widehat{\boldsymbol{Y}} = \boldsymbol{X}_{Sel}\boldsymbol{B}_{\boldsymbol{x}} + \boldsymbol{E}_Y$ (since only few variables are selected)
3. ***Z***$_{orth}$ is obtained by orthogonalizing **Z** with respect to ***X***$_{sel}$
4. Cov-Sel is applied to select variables in ***Z***$_{orth}$
5. **Y** residuals (from step 2) are fitted to ***Z***$_{orth\_sel}$ by OLS: $\widehat{\boldsymbol{E}}_{\boldsymbol{Y}} = \boldsymbol{Z}_{\boldsymbol{orth\_sel}}\boldsymbol{B}_{\boldsymbol{Z}_{\boldsymbol{oorth\_sel}}} + \boldsymbol{E}_{\boldsymbol{Ynew}}$
6. The full model is calculated merging steps 2 and 5: $\widehat{\boldsymbol{Y}} = \boldsymbol{X}_{Sel}\boldsymbol{B}_{\boldsymbol{x}} + \boldsymbol{Z}_{orth\_sel}\boldsymbol{B}_{\boldsymbol{Z}_{\boldsymbol{oorth\_sel}}} + \boldsymbol{E}_{Ynew}$

In case of classification, analogously to PLS-DA the responses matrix is *dummy coded* and the classification rules may be based on predicted responses or on applying LDA on the selected variables ([$\boldsymbol{X}_{sel}$ $\boldsymbol{Z}_{orth\_sel}$]).

The optimal number of variables to be retained in each block can be carried out as explained below for Cov-Sel.

3.5 Variables selection

Variables selection is a research field *per-se* very rich in available methods. The main reasons for aiming at pruning the initial set of collected variables/descriptors to be used in multivariate regression or classification tasks, when using latent variables based methods, are enhancing interpretability, remove extremely noisy variables when these are a huge number, and selecting few informative ones, e.g. to reduce the experimental analyses cost or to build cheaper spectroscopic devices by using only some spectral bands (i.e. using LEDs).

Among the available methods in this Thesis, we evaluated CovSel.

3.5.1 Covariance Selection (CovSel)

CovSel [43] is a variable selection method dedicated to the cases where there is a huge number of variables (yielding a very large solution space), the variables are highly correlated, like in case of spectral signals, and the aim is to obtain very few selected features. CovSel performs variable selection iteratively up to a maximum number decided by the user (falling in the wrapper methods for feature selection). At each selection step the global covariance between single dependent variables with all the responses is evaluated, and the variable showing the highest covariance is selected first; then follow a projection of the data orthogonally to the selected variable before the next selection step. The maximum number of variables to be selected is given in input by the user (i.e. there is not an optimization of a performance criterion to stop the selection) which a posteriori can graphically inspect the explained ***X*** and ***Y*** variance vs. number of selected features (ordered by selection). In addition, or alternatively, the cross validation prediction error as function of number of included selected features can also be evaluated, to decide how many to retain.

CovSel can be applied in exploratory analysis (in this case the selection criterion is based solely on ***X***-variance), in regression, and discrimination tasks.

In regression, ***Y*** consists of continuous responses, and CovSel could be used to make a variable selection based on all responses and then this global selection can be refined for each individual response, e.g. in a second step the ordered selected variables can be evaluated by stepwise addition to see which number will give the minimum cross validation error for each single response.

For discrimination, ***Y*** contains dummy variables codifying class membership, and CovSel is used on this multi-response ***Y***. In this case, to decide the final number of selected variables to retain LDA can be performed on the selected features by stepwise addition (see paragraph 3.4.2).

3.6 Image analysis

The image analysis field is very broad [44], and duly illustration of it is outside the scope of this Thesis. Here, only the basic of the used approaches, and the motivation for using them, are presented.

Two kinds of images have been analysed RGB (i.e. three Vis channels) and hyperspectral acquiring a whole spectrum in the Vis (400- 800 nm) and NIR ranges (800 -2500 nm).

In the RGB case the aim was to detect objects, such as basil stems and leaves from elaboration of basil images taken by the vision system installed in-line (see 2.5.1). However, due to varying illumination conditions a segmentation approach [45] did not gave satisfactory results, then we evaluated pixel-based approaches. One combined wavelet transforms filter (WT) (section

3.7.1), as features enhancement step [46,47], with PLS-DA, for pixel classification; a second one used a deep learning net.

Hyperspectral images (HSI) of pesto were acquired and evaluated with the objective of assessing homogeneity of distribution of the different components. To this aim first, HSI were unfolded then MCR-ALS was applied to resolve the purest components profiles. Finally, to inspect the resolved components distribution features extraction, by different methods, was applied to the refolded concentration matrix (section 3.7.1).

3.6.1 MCR-ALS in image analysis

MCR-ALS could also be used in hyperspectral image analysis [48] to separate the different contributions of the constituent components, i.e. by spectral unmixing, and to study their distribution in the image.

In this case, the HSI (a 3D data array of dimensions pixels_x * pixels_y * wavelengths) is first unfolded pixelwise to obtain a 2D matrix of dimensions pixels_xy (rows) * wavelengths (columns).

On the assumption that each pixel's spectrum is a linear combination of "pure" components spectra, then MCR-ALS can resolve them. When more than one image must be analysed (several samples altogether) a multiset MCR model can be applied by merging the single sample unfolded matrices to obtain a unique matrix of dimensions (samples x pixel_xy) * wavelengths, as shown in Figure 3-4. In this case, a single set of spectral profiles is recovered (same **S** for all samples) and a distinct concentration matrix for each sample ($\mathbf{C}_s$). The latter can be refolded (pixel_x * pixel_y) obtaining a concentration map for each resolved component (i.e. an image showing the spatial distribution of that component), see Figure 3-5 top. Figure 3-5 bottom shows the corresponding "pure components" spectra.

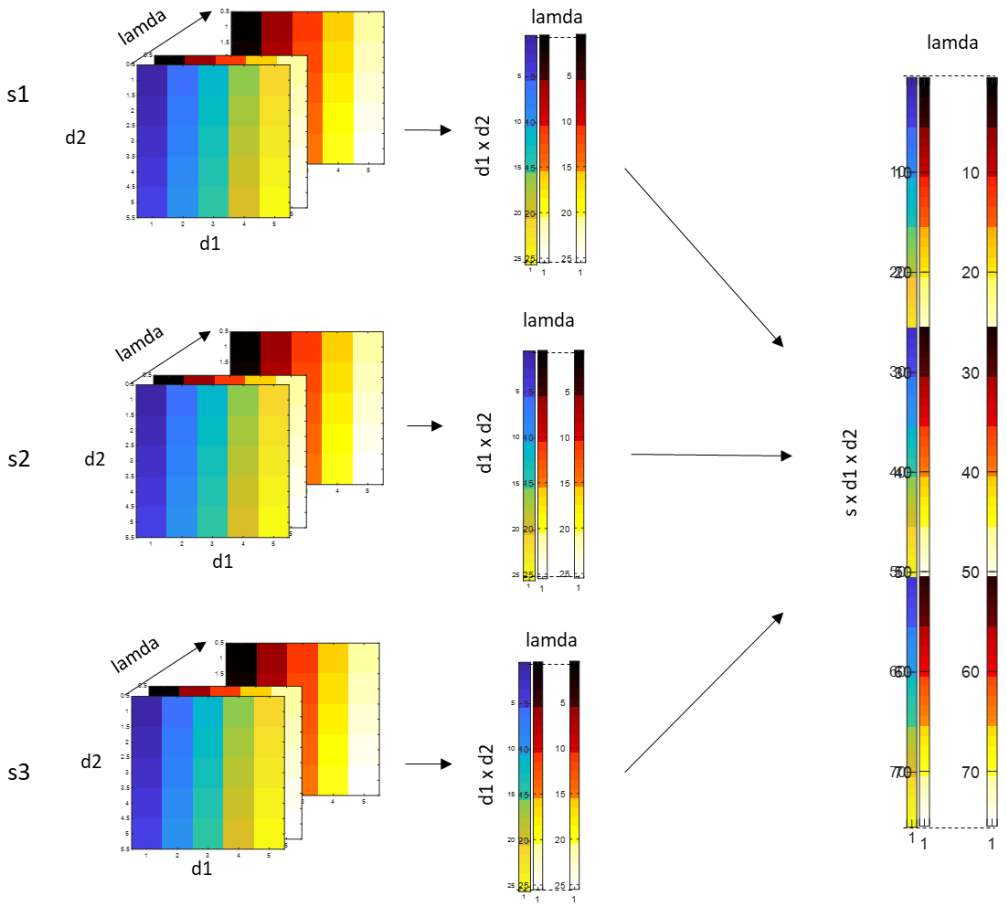


Figure 3-4. Example of 3 unfolded images (corresponding to three distinct samples). HSI arrays are unfolded pixelwise creating 2D matrices of dimensions (d1 x d2) x lambda for each sample s. All the matrices were then merged obtaining a unique matrix of dimension (s x d1 x d2) x lambda that is used in MCR-ALS (multiset option).

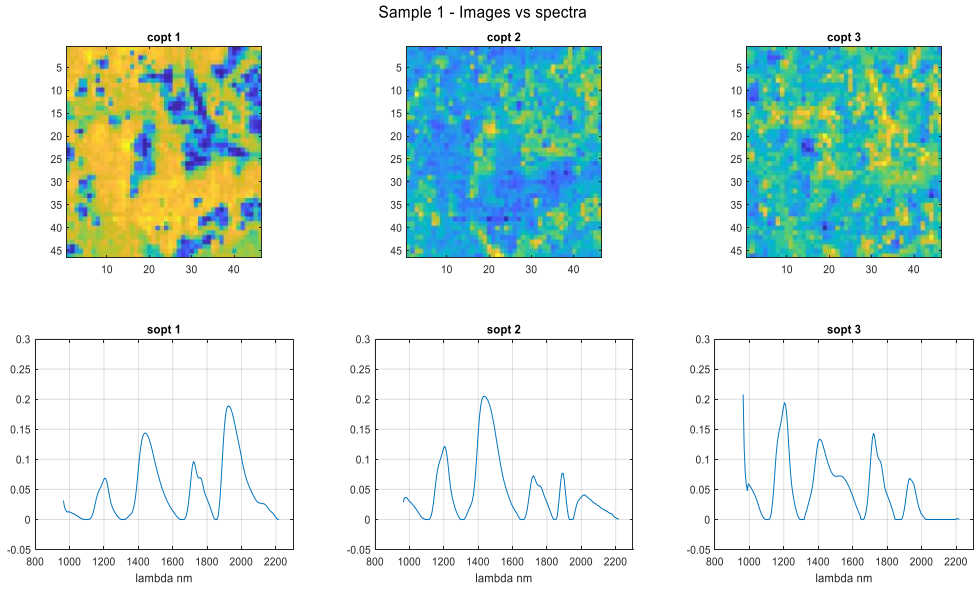


Figure 3-5. Upper part, images of the concentration maps corresponding to each resolved component when analysing by MCR-ALS the NIR hyperspectral image of sample 1. In the lower part the correspondent resolved spectra.

3.6.2 Image features extraction (applied to concentration map)

Different methods have been developed to characterize the spatial distribution of the pixel intensities in images, aiming at obtaining statistical parameters that can differentiate images containing the same elements but in different spatial disposition (texture features) [49].

Texture features extraction methods can be classified in four main categories: (i) statistical, i.e. describing the texture of image regions, by means of high order moments on the pixel frequency histograms; (ii) structural, i.e. defining texture as well-defined compositional elements (spatial regularity of parallel lines); (iii) model based, i.e. which creates an empirical model of the image; and (iv) transform-based, that converts an image in other forms, using filters (e.g. wavelet transform) [50].

In this work, two main approaches were evaluated to assess the concentration map obtained by MCR-ALS (after hyperspectral images decomposition): the well-established Haralick approach [51], and a more recent proposal [52] to evaluate image homogeneity based on comparison of the actual image with one where the same pixels are totally randomly distributed.

3.6.2.1 Haralick features

One of the most used approaches to study image texture (spatial correlation) is the one postulated by Haralick *et al*. in 1973, known as Gray-Level Co-occurrence-Matrix (GLCM) [51]. It consists of two steps: in the first one from the original grayscale image, it is generated a GLCM matrix by considering: one pixel, its grey level, and the level of the surrounding pixels. Each entry (i,j) in a GLCM corresponds to the number of occurrences of the pair of grey levels i and j which are a distance d apart in the original image; in the second step is carried on the calculation of a set of statistical features (angular second moment, contrast, correlation, variance, inverse difference moment, sum average, entropy, energy, etc.) from the GLCM. Haralick proposed 14 statistical features, here only eight of them (which were the less correlated among them) were selected and used (Table 3-1).

Table 3-1. Haralick features formulae. The number of the function in the formula refers to the original Haralick paper.

Feature	**Formula**
Energy (angular second moment)	$$f_1 = \sum_i \sum_j \{p(i,j)\}^2$$ *p(i,j)* is the (i,j)th pixel in image normalized matrix
Contrast	$$f_2 = \sum_{n=0}^{N_{g-1}} n^2 \left\{ \begin{matrix} \sum_{i=1}^{N_g} \sum_{j=1}^{Ng} p(i,j) \\ \lvert i-j \rvert = n \end{matrix} \right\}$$
Correlation	$$f_3 = \frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$ $\mu_x\ \mu_y\ \sigma_x \sigma_y$ are the means and standard deviations of p_x and p_y
Variance	$$f_4 = \sum_i \sum_j (i-\mu)^2 p(i,j)$$
Inverse Different Moment (IDM)	$$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$$
Sum entropy	$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) log\{p_{x+y}(i)\}$$
Information measure of correlation	$$f_{12} = \frac{HXY - HXY1}{max\{HX, HY\}}$$ $$HXY = -\sum_i \sum_j p(i,j) log\ (p(i,j))$$ $$HXY1 = -\sum_i \sum_j p(i,j) log\{p_x(i) p_y(j)\}$$ HX and HY are respectively entropies of p_x and p_y
Maximal Correlation Coefficient	$$f_{14} = (second\ largest\ eigenvalue\ of\ Q)^{1/2}$$ *where* $$Q(i,j) = \sum_k \frac{p(i,k) p(j,k)}{p_x(i) p_y(k)}$$

3.6.2.2 Homogeneity index

This method is based on the Macropixel analysis (MA) introduced by Hamad et al. [53]. MA is a method that splits the image in smaller sub-images and study the different properties of those sub-images and their correlations.

Two possible ways may be used to scan the original image in MA: the Discrete Level tiling (DLT) which uses non overlapping tiles, and the Continuous-Level Moving Block (CLMB) that scans in all possible dimensions the macropixels in the image. CLMB method was used in this work.

For an image of dimensions L x L (squared for simplicity of explanation), considering a S_m windows of m x m (where m ≤ L) pixels, the total number of sub-images will be:

$$TOTAL_{Sm} = (L-(m-1))\ (L-(m-1)) \qquad Eq.\ 3\text{-}16$$

With a sub-image (S_m) dimension of

$$PIX_m = m * m \qquad Eq.\ 3\text{-}17$$

For each S_m used, the pooled standard deviation is calculated as

$$STD_{Sm} = \sqrt{\frac{\sum\sum(TOTAL_{Sm}-\bar{s})}{PIX_m-1}} \qquad Eq.\ 3\text{-}18$$

Where $\bar{s}$ is the average of the pixel intensity of the whole image. The standard deviation of each sub-sample windows m will be:

$$S_{wm} = \frac{\sum STD_{Sm}}{TOTAL_{Sm}} \qquad Eq.\ 3\text{-}19$$

Plotting the S_{wm} vs r, the normalized windows dimension (pixel size/image pixel size) a so-called homogeneity curve is obtained.

In other words, we can say that homogeneity curve is the results of the application of CLMB analysis to evaluate the mean standard deviation of the macro-pixels in an image.

Homogeneity curves were calculated with an algorithm in MATLAB supplied by courtesy of Prof. Jose Amigo, University of Basque Country (Spain).

The results provide a comparison of the homogeneity calculated for: (i) the actual image; (ii) the Homo image, calculated on the randomized image (where the pixels intensities are the same, as in the original one, but uniformly distributed); (iii) the Inhomo image, calculated on an image obtained by arranging the pixels in an ordered pattern.

Inhomo image represents the maximum level of possible inhomogeneity for the studied image, while the Homo image the maximum possible homogeneity level. The algorithm calculates the relative homogeneity (%H) of the image, as percentage of the difference of homogeneity between the actual and the Inhomo, divided by the difference between the Homo and Inhomo.

An example of results is reported in Figure 3-6.

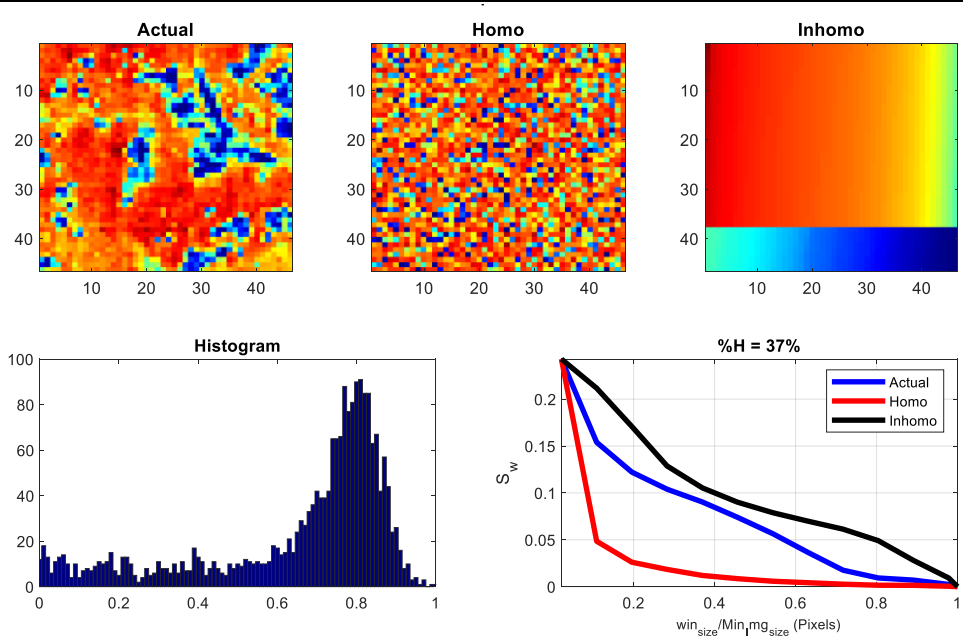


Figure 3-6. Example of Homogeneity results.

3.7 Analysis of RGB images

3.7.1 Wavelet (WT) + PLS-DA approach

Wavelet transform (WT) is a very powerful mathematical tool used to extract information form many kinds of data, included images [54].

Wavelet is categorized into continuous wavelet tools and discrete wavelet tools [55]. The first are used for signal analysis or time-frequency analysis, while the seconds are most used for compressing data [56]

The basic idea of the wavelet transform is to decompose a signal or an image into distinct subspaces capturing different frequency contents of the raw signal/images, namely high frequencies are collected in the so-called details blocks (holding sharp, oriented changes, etc) while low frequencies in approximation (holding smooth changes like tones) block.

In the case of images, to do that 2D-WT applies recursively high and lowpass filters to obtain four sub-images: 1) approximation (A): a low-pass filter is applied both row- and column-wise; 2) horizontal details (H): a low-pass filter is applied row-wise, then a high-pass filter, column-wise; 3) vertical details (V): a high-pass filter is applied row-wise, then a low-pass filter, column-wise; 4) diagonal details (D): a high-pass filter is applied both row- and column-wise.

This decomposition can be then applied to the obtained approximation block, obtaining A, H, V and D at second decomposition level, and so on until the maximum decomposition level compatible with the image size is reached.

This decomposition is applied distinctly to each spectral channel.

The four sub-images for each channel are then unfolded pixel-wise and concatenated to obtain a final matrix of dimensions: (pixel x pixel), on rows, and (n° of channels x n° of levels x 4) on columns dimension.

This matrix is used in a PLS-DA model where the ***Y*** is a dummy matrix with class membership of each pixel, e.g. 1 in the pixels representing the characteristic of interest to predict in the image (i.e. the background) and 0 in the other pixels.

The obtained PLS-DA model is then applied on new images to predict the class to which their pixels belong (i.e. showing the spatial features of the new images), applying the same sequence of steps: wavelet decomposition and unfolding. More details are reported in the chapter 6.

3.7.2 DeepL approach

For the application of the deep learning approach [57,58,59,60] three main steps must be done: in the first, the images have been split in sub images of size 256x256xRGB; then convolution filters have been applied to extract features containing relevant information in the learning phase. The extracted features have been used to recreate pixelwise label space.

The architecture giving the best classification results, among the three tested, consists of a two-layer CNN (convolutional neural networks), where the first layer is the largest (32 units), and the second layer serves to condense the information. These first two convolutional layers encode the information, and the transposed convolution operation serves to decode the information going back to pixel space. The last convolution layer uses the decoded information to learn classification.

3.8 SOFTWARE

Data elaboration was performed within MATLAB (The Mathworks Inc., Natick, MA, USA, 2007) environment. PLS Toolbox 9.1 (Eigenvector Research, Inc., Manson, WA, USA) has been used for PCA, PLS and PLS-DA. The MATLAB Wavelet Toolbox has been used for image decomposition, while the Image Processing MATALAB Toolbox has been used for GLCM image analysis.

For MCR-ALS the MCR-ALS GUI 2.0 [61] has been used, which can be freely downloaded from the website www.mcrals.info.

The CovSel code (in Matlab) has been implemented and kindly provided by courtesy of Prof. Jean Michel Roger (French National Institute for Agriculture, Food, and Environment (INRAE).

The Homogeneity code (in Matlab) performing Continuous Level Moving Block method has been implemented and kindly provided by courtesy of Prof. José Amigo (University of Basque Country, Spain).

The SO-PLS and SO-PLS-DA codes (Matlab) have been implemented and kindly provided by the Rome Chemometrics group (Prof. Federico Marini and Prof. Alessandra Biancolillo) of University La Sapienza (Rome, Italy).

Several auxiliary routines have been implemented in Matlab by me or by the Modena chemometrics research group.

---

1 Wold S. Chemometrics; what do we mean with it, and what do we want from it?, Chemom. Intell. Lab. Syst. 1995; 30: 109–115. doi: 10.1016/0169-7439(95)00042-9.

2 Comprehensive Chemometrics, Chemical and Biochemical Data Analysis. 2nd Edition - May 26, 2020. Editors: Steven Brown, Roma Tauler, Beata Walczak

3 Handbook of Chemometrics and Qualimetrics: Part A By Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, and Smeyers-Verbeke J. Data Handling in Science and Technology Volume 20A. Elsevier: Amsterdam. 1997

4 Brereton RG. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. John Wiley & Sons, Ltd., Chirchester, England, 2003

5 Brereton RG, Jansen J, Lopes J et al. Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. Anal Bioanal Chem. 2017; 409: 5891–5899. https://doi.org/10.1007/s00216-017-0517-1

6 Marini F. (Ed.) (2013). Chemometrics in food chemistry. Newnes.
In Data Handling in Science and Technology, Vol. 28, published by Elsevier.

7 Kumar N, Bansal A, Sarma GS, Rawal RK, Chemometrics tools used in analytical chemistry: An overview, Talanta. 2014; 123:186-199 https://doi.org/10.1016/j.talanta.2014.02.003.

8 Roberts JJ, Cozzolino D. An overview on the application of chemometrics in food science and technology—An approach to quantitative data analysis. Food Analytical Methods, 2016; 9(12): 3258-3267.

*9 Cocchi M. Chemometrics for food quality control and authentication, in Encyclopedia of Analytical Chemistry R.A. Meyers (Ed.), John Wiley & Sons Ltd, **2017,** pp. 1-29. 10.1002/9780470027318.a9579*

10 Cocchi M, Li Vigni M, Durante C. Chpt. 17 Chemometrics Bioinformatic. In C. A. Georgios, P. Danezis (Ed.) Food Authentication: Management, Analysis and Regulation, J. Wiley & sons, 2017, pp. 483-520. doi: 10.1002/9781118810224.ch17. ISBN13 9781118810262.

11 Trygg J. Data Preprocessing, in Comprehensive Chemo- metrics, eds S.D. Brown, R. Tauler, B. Walczak, Elsevier B.V., 1–129, Vol. 2, 2009.

12 Engel J, Gerretzen J, Szymanska E, Jansen JJ, Downey G, Blanchet L et al. Breaking with Trends in Pre-Processing? TrAC Trends in Anal. Chem. 2013; 50: 96–106.

13 Rinnan A, van den Berg F, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Anal. Chem. 2009; 28(10): 1201-1222.https://doi.org/10.1016/j.trac.2009.07.007.

14 Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. TrAC Trends in Analytical Chemistry, 132, 116045.

15 Bro R, Smilde AK. Centering and scaling in component analysis. J Chemometr.2003; 17: 16–33.

16 Savorani F, Tomasi G, Engelsen SB. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. J. Magn. Reson. 2010; 202: 190-202 doi: 10.1016/j.jmr.2009.11.012

17 Tomasi G, Savorani F, Engelsen SB. icoshift: An effective tool for the alignment of chromatographic data. J Chromatogr A. 2011;1218(43): 7832-7840. doi: 10.1016/j.chroma.2011.08.086. Epub 2011 Sep 3. PMID: 21930276.

18 Wise BM, Gallagher NB, Bro R, Shaver JM, Windig W, Chemometrics tutorial for PLS_Toolbox and Solo - Eigenvector Research, Inc, 2006

19 Proceedings of the 1963 International Symposium on Humidity and Moisture, Principles and Methods of Measuring Moisture in Liquid and Solids, vol. 4, Reinhold Publishing Co., New York, 1965, p.19

20 Burns DA, Ciurzak EW, editors. Handbook of near infrared analysis. 3rd Edition. New York: CRC Press Taylor & Francio Group, 2008.

21 Pasquini C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. 2003; 14(2): 198-219.

22 Pasquini C. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review, Analytica Chimica Acta. 2018; 1026: 8-36. https://doi.org/10.1016/j.aca.2018.04.004.

23 Li Vigni M, Durante C, Cocchi M. Chpt. 3 Exploratory Data Analysis. In F. Marini (Ed.) Chemometrics in Food Chemistry Vol. 28 Data Handling in Science and Thecnology series, Elsevier 2013, ISBN 9780444595287. doi: 10.1016/B978-0-444-59528-7.00003-X

24 Jolliffe IT. Principal Component Analysis,2nd Edition, in Springer Series in Statistics Springer New York, NY, 2002 https://doi.org/10.1007/b98835

25 Bro R, Smilde AK, Principal Component Analysis, Anal. Methods. 2014: 6: 2812 doi: 10.1039/c3ay41907j

26 Ringér M. What is principal component analysis? Nature Biotechnology 2008;, 26: 303-304https://doi.org/10.1038/nbt0308-303

27 Cattell RB. The Scree Test For The Number Of Factors, Multivariate Behavioral Research.1966; 1(2):, 245-276. doi: 10.1207/s15327906mbr0102_10

28 de Juan A, Tauler R. Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review. Analytica Chimica Acta. 2021;1145: 59-78. https://doi.org/10.1016/j.aca.2020.10.051.

29 Ruckebusch C. Blanchet L. Multivariate curve resolution: A review of advanced and tailored applications and challenges. Analytica Chimica Acta. 2013; 765: 28-36.https://doi.org/10.1016/j.aca.2012.12.028.

30 de Juan, Anna, Joaquim Jaumot, and Romà Tauler. "Multivariate Curve Resolution (MCR). Solving the mixture analysis problem." Analytical Methods 6.14 (2014): 4964-4976.

31 Stevens, J. P. (2002). Applied multivariate statistics for the social sciences. Mahwah, NJ: Lawrence Erblaum.

32 Bertinetto C, Enge J, Jansen J, ANOVA simultaneous component analysis: A tutorial review, Analytica Chimica Acta. 2020; X 6:100061. https://doi.org/10.1016/j.acax.2020.100061.

33 Måge, I. and Marini, F. (2023), Advancements in multivariate analysis of variance. Journal of Chemometrics, 37: e3504. https://doi.org/10.1002/cem.3504

34 Svante Wold, Michael Sjöström, Lennart Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, Volume 58, Issue 2, 2001, Pages 109-130, ISSN 0169-7439, https://doi.org/10.1016/S0169-7439(01)00155-1.

35 Wold, H. Nonlinear estimation by iterative least square procedures. In: David, F.N. (ed.): Research papers in statistics, Festschrift for J. Neyman, Wiley, New York, 411-444, 1966.

36 de Jong, S., SIMPLS: an alternative approach to partial least squares regression, Chemometrics and Intelligent Laboratory Systems, 1993, 18, 251-263.

37 Stahle L, Wold L, Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study, J. Chemometr. 1987; 1: 185–196.

38 Barker M, and Rayens W. Partial least squares for discrimination. Journal of Chemometrics: A Journal of the Chemometrics Society. 2003; 17(3): 166-173.

39 Indahl UG, Martens H, Næs T., From Dummy Regression to Prior Probabilities in PLS-DA Extraction and Classification. J. Chemom. 2007; 21: 529–536.

40 Tang L, Peng S, Bi Y, Shan P, Hu X. A New Method Combining LDA and PLS for Dimension Reduction. PLoS ONE. 2014; 9(5): e96944. https://doi.org/10.1371/journal.pone.0096944

41 Biancolillo, A., & Næs, T. (2019). The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions. In Data handling in science and technology (Vol. 31, pp. 157-177). Elsevier.

42 Næs T, Romano R, Tomic O, Måge I, Smilde A, and Liland K H. Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. Journal of Chemometrics, 2021; 35(10): e3243.

43 Roger JM, Palagos B, Bertrand D, Ferrandez-Ahumada E. CovSel: Variable selection for highly multivariate and multi-response calibration Application to IR spectroscopy. Chemometrics and Intelligent Laboratory Systems. 2011; 106: 216–223

44 Jose Amigo (Ed.) Hyperspectral imaging. Sep 2019 Elsevier

45 Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics. 1979; 9(1): 62–66.

46 Prats-Montalban JM, Cocchi M, and Ferrer A. Nway modeling for wavelet filter determination in multivariate image analysis. Journal of Chemometrics. 2015; 29(6): 379‑388

47 Li Vigni M, Prats-Montalaban JM, Ferrer A, Cocchi M. Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA) J. Chemometrics. 2018; 32, e2970-e2990.

48 Chapter 2.5 - Multivariate curve resolution for hyperspectral image analysis, Editor(s): José Manuel Amigo, Data Handling in Science and Technology, Elsevier, Volume 32, 2019, Pages 115-150, ISSN 0922-3487, ISBN 9780444639776, https://doi.org/10.1016/B978-0-444-63977-6.00007-9.

49 Aptoula E, Lefevre S. Morphological texture description of grey-scale and color images, in: P. Hawkes (Ed.), Adv. Imaging Electron Phys, Academic Press, 2011, pp. 1–74, https://doi.org/10.1016/B978-0-12-385981-5.00001-X.

50 Bharati MH, Liu JJ, MacGregor JF. Image texture analysis: methods and comparisons, Chemom. Intell. Lab. Syst.2004; 72: 57–71. https://doi.org/10.1016/j.chemolab.2004.02.005.

51 Haralick RM, Shanmugam K, and Dinstein I. Textural Features for Image Classification., IEEE Trans. on Systems, Man, and Cybernetics. 1973; SMC-3(6): 610-621.

52 de Moura França L, Amigo JM, Cairós C, Bautista M, Pimentel MF. Evaluation and assessment of homogeneity in images. Part 1: Unique homogeneity percentage for binary images, Chemometrics and Intelligent Laboratory Systems. 2017;171: 26-39. https://doi.org/10.1016/j.chemolab.2017.10.002

53 Hamad ML, Ellison CD, Khan MA, Lyon RC. Drug product characterization by macropixel analysis of chemical images, J. Pharm. Sci. 2007; 96: 3390–3401. https://doi.org/10.1002/jps.

54 Jaideva GC and Chan AK. Fundamentals of wavelets: theory, algorithms, and applications. John Wiley & Sons, 2011.

55 Hoang VD. Wavelet-based spectral analysis, TrAC Trends in Analytical Chemistry. 2014;62: 144-153 https://doi.org/10.1016/j.trac.2014.07.010.

56 Beata Walczak. Wavelets in chemistry. Elsevier eBooks, 1 2000.

57 Alzubaidi L, Zhang J, Humaidi AJ et al Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021; 8: 53 https://doi.org/10.1186/s40537-021-00444-8

58 Li Y, Zhang H, Xue X, Jiang Y, Shen Q. Deep learning for remote sensing image classification: A survey. WIREs Data Mining Knowl Discov. 2018; 8: e1264. https://doi.org/10.1002/widm.1264

59 Russel NS, Selvaraj A. Leaf species and disease classification using multiscale parallel deep CNN architecture. Neural Comput & Applic. 2022; 34: 19217–19237. https://doi.org/10.1007/s00521-022-07521-w

60 Lguensat R, Sun M, Fablet R, Tandeo P, Mason E and Chen G. EddyNet: A Deep Neural Network For Pixel-Wise Classification of Oceanic Eddies. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain 2018; 1764-1767. doi: 10.1109/IGARSS.2018.8518411.

61 Jaumot J, de Juan A, Tauler R, MCR-ALS GUI 2.0: New features and applications, Chemometrics and Intelligent Laboratory Systems. 2015; 140: 1-12. https://doi.org/10.1016/j.chemolab.2014.10.003.

4BASIL AROMA CHARACTERISATION

Basil aroma is one of the main traits that confers ageable sensory features to the pesto sauce. Its characterization is of utmost importance, both in the search for new basil chemotypes (also called varieties in common language and in the paper 1) and in the control of the basil used. To this aim a fast and sensible analytical technique (GC-FID based e-nose) was used and two different analysis approaches were evaluated: i) target analysis (implying use of analytical standards and quantification), which required prior to GC-FID e-nose analysis the use of reference techniques (GC-MS and olfactometry) to assess the key odorant molecules present and perceivable; and ii) untargeted analysis, *i.e.* direct application of GC-FID e-nose and elaboration of the whole chromatographic profile. Untargeted approach is advantageous in terms of analysis cost/time and for recovering the entire information since the whole aroma fingerprint is considered. However, it requires proper chemometric tools. Within my Thesis objective this is an example of how a deeper chemometric knowledge in R&D may aid developing faster approaches in routine analysis.

4.1 Targeted analysis of basil aroma

Here the study context and results are summarized, for more details, please refer to published paper number 1 in appendix 1.

The basil aroma is composed of many molecules, mainly terpenoids, alcohols, aldehydes, ketones, and esters [1,2]. Totally, there are more than one hundred molecules, of which the most representatives in sweet basil are considered linalool, estragole, eugenol and eucalyptol (1,8-cineole) [3,4]. The content of these molecules could give a preliminary evaluation of different basil flavour profiles, while a more accurate evaluation of the final aroma will also consider the concentrations of other minor components, mainly the molecules that have a low odour threshold [5,6]. The odour threshold is defined as the lowest concentration of a molecule that could be perceived by olfaction. Thus, in the evaluation of the flavour patterns, it is necessary to consider not only the concentration of a given molecule but also its capacity to be perceived.

Despite there are many different methods to identify and quantify volatile organic compounds (VOCs), the basil aroma pattern, to the best of my knowledge, has been characterized only by gas chromatography (GC) based techniques like for instance, headspace solid phase microextraction gas chromatography–mass spectrometry (HS-SPME-GC–MS) [7], headspace sorptive ex-traction gas chromatography–mass spectrometry (HSSE GC–MS) [8], as well as gas chromatography as such (GC and GC–MS) [4] indirectly measuring the total phenolic compounds [9] or using flow-injection mass spectrometry [10].

As basil is a very delicate plant, which is difficult to store after cutting [11,12], it would be extremely useful to have a fast analytical method, being at the same time suitable to discriminate the different chemotypes and furnishing information on the compositional profile of the aroma fraction.

To this aim, in my work an GC-FID electronic nose device (Heracles II, Alpha MOS, Toulouse, France) was tested since it can provide a rapid and sensitive system.

The basil key odorant molecules were selected combining information from sensory evaluation and gas-chromatography olfactometry (see published paper number 1) Then the nine key molecules individuated were quantified, in a fast way by using GC-FID e-nose and calibration by external standards, with an internal standard to normalize every single injection.

Several basil chemotypes were analysed, grown on open fields in different years and considering more cuts each year. The aim was obtaining a preliminary over-view by multivariate

exploratory data analysis of the aroma variation due to both chemotypes and period of harvesting. To deepen understanding of these effects and to assess their statistical significance ANOVA–Simultaneous Component Analysis (ASCA) was used [13]. ASCA generalizes classical analysis of variance (ANOVA) to multivariate data, over-coming the main limitations (number of samples higher than number of variables, breakdown in case of variables collinearity) and multinormal distribution assumption of multivariate ANOVA (MANOVA).

First, a classic ANOVA was carried out to split the data matrix into the effect matrices for each experimental factor and their interactions. Then, simultaneous component analysis was carried out on the effect matrices to identify and visualize the contribution of the measured variables to each of the effects that introduced systematic variation [14].

Because ASCA requires data coming from an experimental design, and sampling was not programmed beforehand having ANOVA analysis in mind, a balanced reduced set (to meet a balanced design) of basil samples was selected, to investigate the effects of cutting period (cut), basil chemotypes and harvesting year on the basil aroma pattern.

4.1.1 Results and Discussion

4.1.1.1 Basil aroma analysis for molecules identification

The pattern of volatile compounds of basil highlighted by the fast-CG analysis comprises eighteen molecules that were tentatively identified by using the Kovats relative retention indexes. The Heracles software compares the retention indexes of the two columns which have different polarities to improve the tentative identification. In Figure 4-1, the identified molecules are shown. Among them, there are the nine ones that were identified as relevant in terms of persistent perceived odour by applying olfactometry analysis (GC-O) with an expert panel. Thus, this is an indication that the fast-CG technique is suitable to characterise basil aroma.

The identification of these nine molecules was confirmed by comparison with the elution time of injected standards, once peaks were identified, calibration curves for quantification were obtained by using an internal standard. The resulting concentration values were consistent with a typical “eucalypt” basil volatile pattern [6,8] with the prevalence of linalool, followed by eucalyptol (1,8-cineole) and then by eugenol. Other molecules are typical of essential oils of basil such as hexanal, α-pinene, myrcene and β-caryophyllene [12].

As previously reported, the flavour profile is strictly related to the presence or the prevalence of key odorant molecules, with a consequent impact on the final perceived bouquet. Four main basil chemotypes have been described by Lawrence et al. [15] depending on the prevalence of odorant molecules: estragole rich, linalool rich, methyl-eugenol rich and methyl cinnamate rich. Chemotypes used in the present study held predominantly in the linalool rich chemotype, but with some diversity. Chemotype 8, for example, was characterized for its lower level of linalool compared to other varieties, whereas on the contrary, chemotype 9 had the highest content. In a similar way, estragole was relatively more present in chemotypes 8 and 9 with respect to other chemotypes.

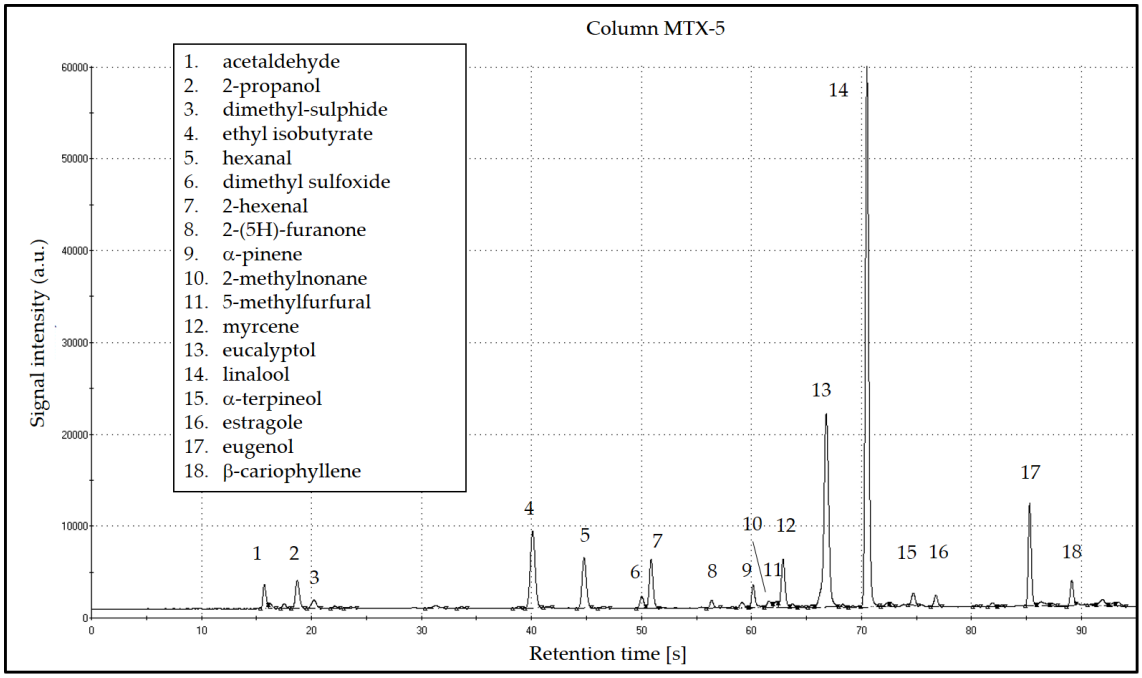


Figure 4-1. As example it is shown the chromatogram obtained by elution on column MXT-5 of Heracles II for one of the samples. Peak 4 is the internal standard.

4.1.1.2 Multivariate Exploratory Data Analysis

PCA analysis was applied to the autoscaled data matrix composed by the nine volatile molecules (variables, in column) obtained for the 267 samples (rows) characterized by different varieties, cuts, and harvested years. Autoscaling was selected as the most appropriate data pre-processing method as the different volatile compounds had different variances due to their different concentration ranges.

In this first exploratory analysis, two principal components seemed appropriate considering their explained variance (Figure 4-2).

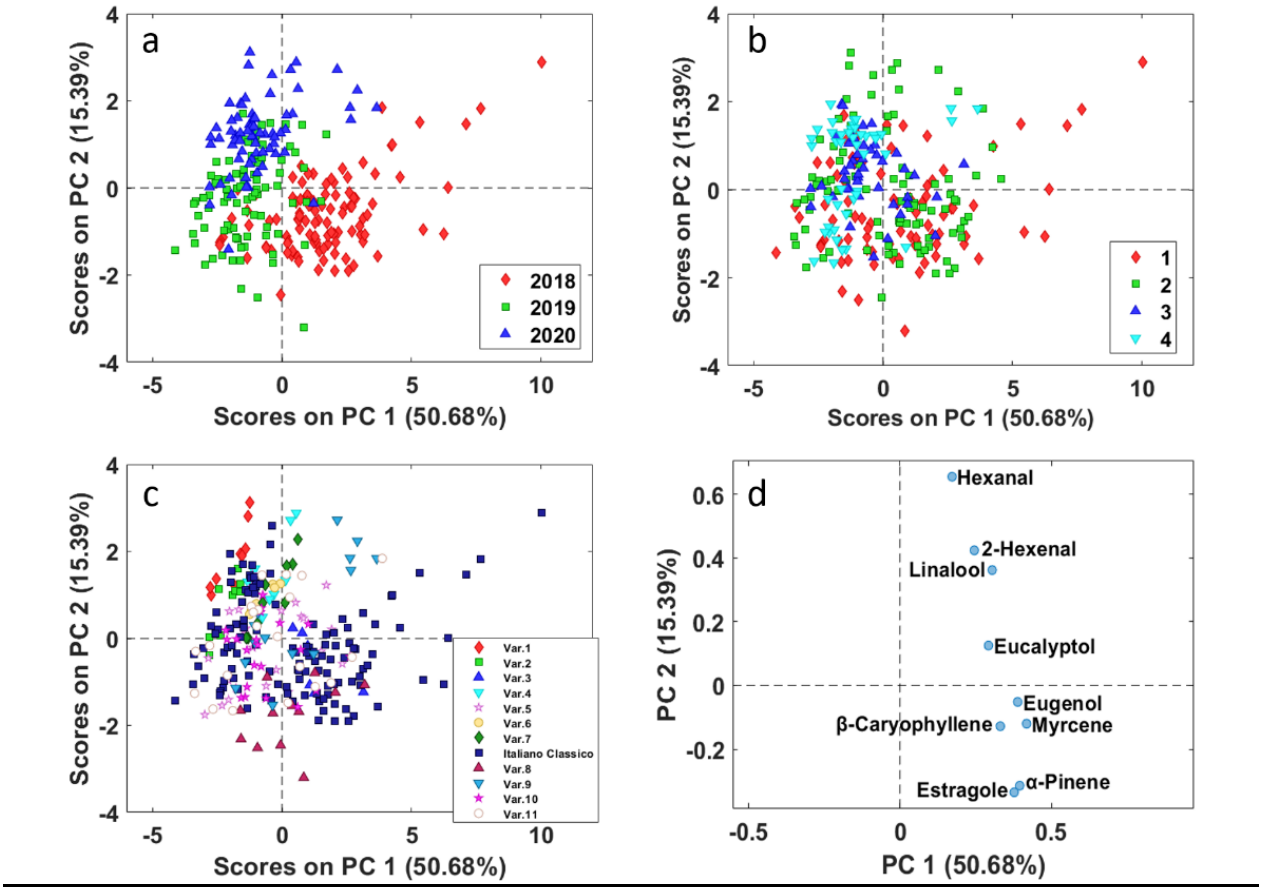


Figure 4-2. PCA of all basil samples. PC1 vs. PC2 scores (**a**–**c**) and loadings (**d**) plots. Basil samples are coloured according to: (**a**) year; (**b**) cut; (**c**) chemotype.

In Figure 4-2, the PC1 vs PC2 scores plot (PCA conducted on species concentrations) is reported and the different basil samples are represented with different symbols and colour according to year (Figure 4-2a), cut (Figure 4-2b) and basil chemotype (Figure 4-2c).

From the PCA results some information could be obtained. Figure 4-2a shows that slight differences could be observed among the three harvesting years, more in 2018 than in 2019 and 2020. The main contribution to this separation seems to be due to a higher concentration of almost all the investigated volatile molecules, since they lie on the same side of the respective loadings plot, all at positive values (Figure 4-2d). This difference is within the expected yearly variability, due to the different weather conditions. As an example, the year 2018 was probably characterized by less rainfall than the years 2019 and 2020.

As far as different basil cuts are concerned, Figure 4-2b points out that well defined clusters are not observable with respect to different basil cuts. Cut number 4, located on the left of the scores plot, is more homogeneous, at first it seems that the average level of all the flavour molecules is lower than in the other cuts; however, this information overlaps with that of the year.

In Figure 4-2c, the different chemotypes are rather overlapped, and it is evident a “spread” of “Italiano Classico” basil chemotype samples, which are uniformly distributed along the variability range of the scores space. Notwithstanding, PC2 highlights the difference of basil chemotype 8, which has the most negative scores on PC2 and thus presents a higher value of estragole and α-pinene (negative loadings values on PC2). A few samples harvested in 2020 of chemotypes 1, 4 and 9, and of “Italiano Classico” harvested in 2018, show high positive scores value on PC2, corresponding to higher amount of hexanal (most positive loadings value on PC2), whose odour is described as “green grass”, and could give, depending on its concentration, an unwanted “hay” note.

Finally, it can be observed that chemotypes 1, 2, 4, 6 and 7, which were cultivated only in 2020, are mostly located in the first quadrant (negative PC1 and positive PC2 score values) this indicates a lower amount of estragole, α-pinene, myrcene, β-caryophyllene, and eugenol, which

fall in the opposite quadrant in the loadings space (positive PC1 and negative PC2 loading values) and thus less fruity/floral and spicy odours.

In general, the interpretation of the overall PCA results is hampered due to the combined effects of all the investigated factors.

For these reasons ASCA was applied on the balanced reduced dataset with the aim to assess if the considered experimental factors and their interactions could have a significant effect on basil's aromatic profile. The effects/interactions partition by ASCA is reported in Table 4-1 (first column) together with the significance (p-value, second column) of each term effect as assessed by means of a permutation test (i.e. by comparing the experimental sum of squares for each effect matrix with its corresponding distribution under the null hypothesis).. All the considered factors and interactions were statistically significant ($p < 0.05$), even though the effects of the factors "chemotype" and "year" presented a higher explained variance than other effects. On the other hand, the effect of factor "cut" explained just 3% of the total variance, suggesting a lower influence on basil's aromatic profile compared with the other two main factors. This can also be seen in the fact that the second order interactions in which factor "cut" is involved explain less than the 4% of the total variance, whereas the interaction "year × chemotype" explains about the 12%.

Table 4-1 Explained variance and probability values for main factors and their second order interactions.

Factor	Explained Variance %	*p*
Chemotype	36.41	<0.001
Year	22.31	<0.001
Year × Chemotype	11.95	<0.001
Year × Cut	3.74	<0.001
Cut × Chemotype	3.1	0.003
Cut	3	<0.001

After the assessment of the significance of each factor and interaction, a component analysis (SCA) was performed on each effect matrix separately to interpret the observed variation. In Figure 4-3a, the scores plot of the effect for factor "year", with projected residuals, is shown. Since the year effect matrix contains just two rows, one for each considered year, the SCA model is described by only one component (SC1), which explains 100% of the variance.

From the scores plot, it was possible to confirm the significant difference between the two levels of the factor "year": all samples collected in 2019 have negative scores, whereas almost all the samples collected in 2020 have positive scores, highlighting the high difference between the two levels of this factor. To explain this difference, in Figure 4-3b the corresponding loadings plot is reported, where it can be observed that the year 2020 samples present higher contents of almost all the molecules investigated in the study, except for 2-hexenal and myrcene, which do not contribute to explain the difference between the two years.

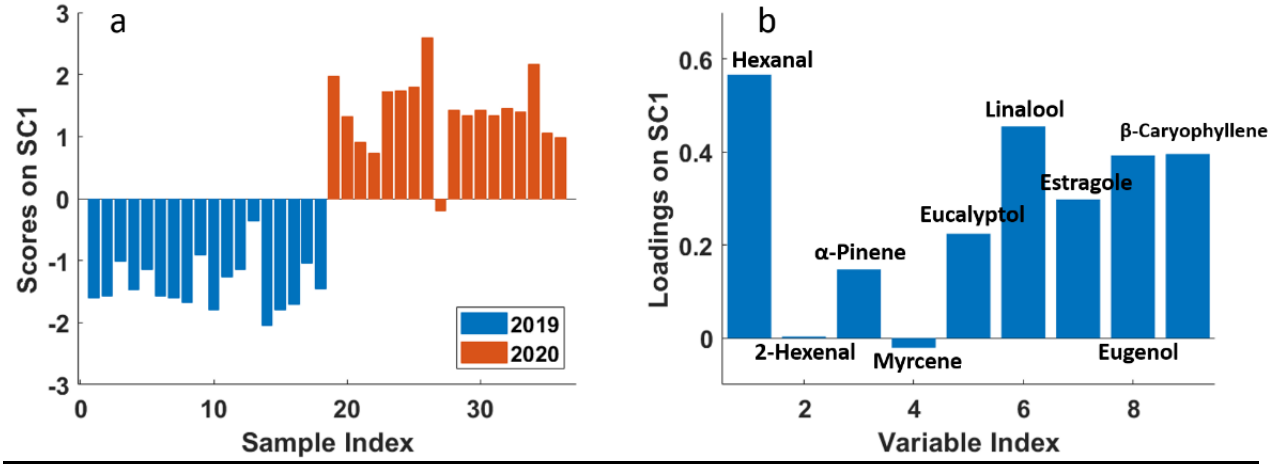


Figure 4-3. SCA of the effect matrix "year". (**a**) Scores plot (SC1) with projected residuals; (**b**) variable loadings (SC1).

Figure 4-4 a,b shows the scores and loadings plots for the effect of factor "cut", respectively. They are represented in the same way as for the factor "year". In this case, the scores plot

confirms that there is a significant difference between the second and fourth cuts, even if it is not as marked as for the other main factors. Scores of samples from 10 to 18 (4th cut, year 2019) present both positive and negative values in an irregular pattern. From the loadings plot, it is possible to observe that samples collected at the fourth cut present mainly a higher content of myrcene, eugenol and linalool, with respect to the second cut samples. β-caryophyllene and 2-hexenal contribute to the same direction but to a lesser extent. A slightly lower content of estragole characterizes the second cut. In general, for the factor "cut", not all the samples characterized by the same conditions behave similarly, as the effect of "cut" is of the same entity of its interactions with year and chemotype. However, the general trend suggests that the influence of this factor on basil's aromatic profile cannot be neglected.

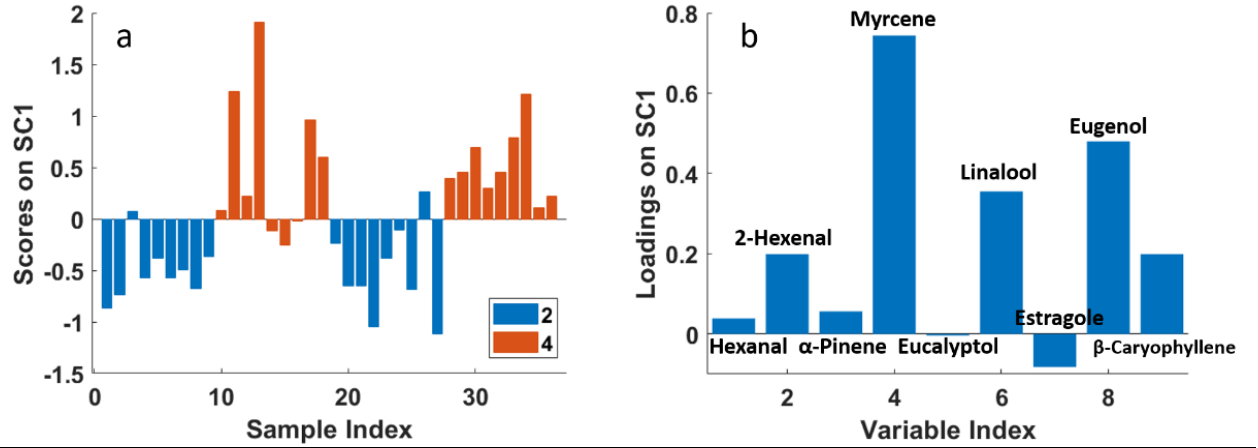


Figure 4-4. SCA of the effect matrix "cut". (**a**) Scores plot (SC1); (**b**) variable loadings (SC1).

Results of SCA for the factor "chemotype" are represented in Figure 4-5. In this case, since the factor "chemotype" was varied at three levels, two components (SCs) were necessary to describe its effect. The first SC clearly describes the difference between Var. 9 with respect to Var. 5 and "Italiano Classico" chemotypes. Var. 9 presented a higher content of almost all the molecules considered in this study, especially eucalyptol, estragole, and α-pinene, which gave a balsamic connotation to the odour. On the other hand, the second SC shows the difference between Var. 5 and "Italiano Classico" chemotypes, less marked than the difference described by SC1. In this case, the compounds mainly responsible for this difference are hexanal and 2-hexenal, which are in greater quantity in the "Italiano Classico" chemotype, whereas Var. 5 is characterized by slightly higher quantities of eugenol, β-caryophyllene, α-pinene, estragole and eucalyptol.

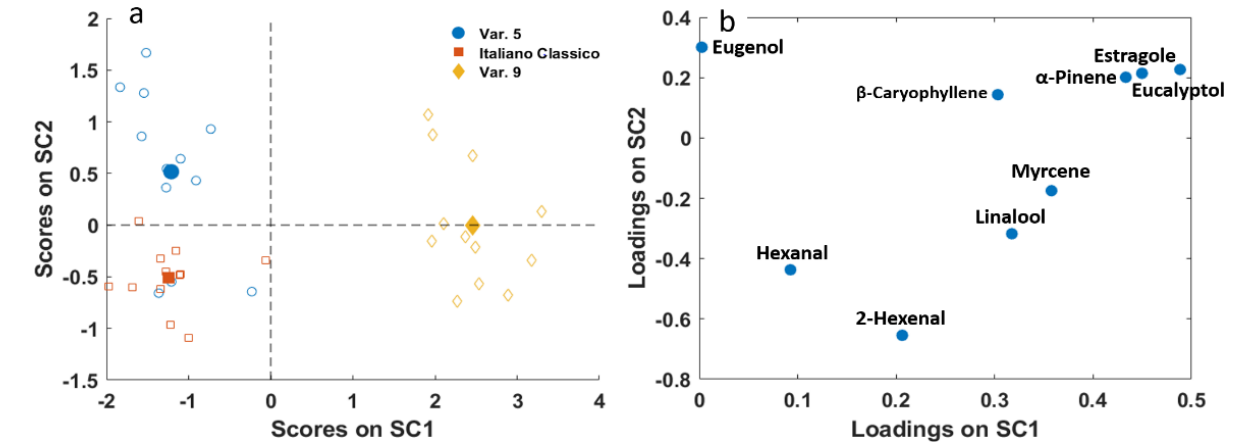


Figure 4-5. SCA of the effect matrix "chemotype". (**a**) SC1 vs. SC2 scores plot with projected residuals (empty symbols); (**b**) variable loadings (SC1 vs. SC2).

To deeply investigate the effect of considered factors on basil's aromatic profile, their second order interactions were also examined. Figure 4-6 shows the effect of the interaction between the

factor's "year" and "chemotype". It is possible to observe how Var. 9 is extremely different from the other two chemotypes, as it shows the opposite behaviour in SC1, i.e., Var. 9 samples collected in 2020 (negative SC1 values) have a higher content of almost all the considered molecules (negative SC1 loadings, except for 2-hexenal and hexanal close to zero) with respect to samples of the same chemotype collected in 2019. At variance, the other two chemotypes are richer in flavours in 2019 than in 2020. "Italiano Classico" and Var. 5 show the opposite behaviour with respect to year in SC2: the first is richer in flower/fruity aroma (higher myrcene and linalool) and lower in α-pinene and hexanal in 2019 with respect to 2020, and the opposite holds for Var. 5. Thus, it is worth noting how the variation of the factor "year" changes the chemical composition of samples of the same chemotype.

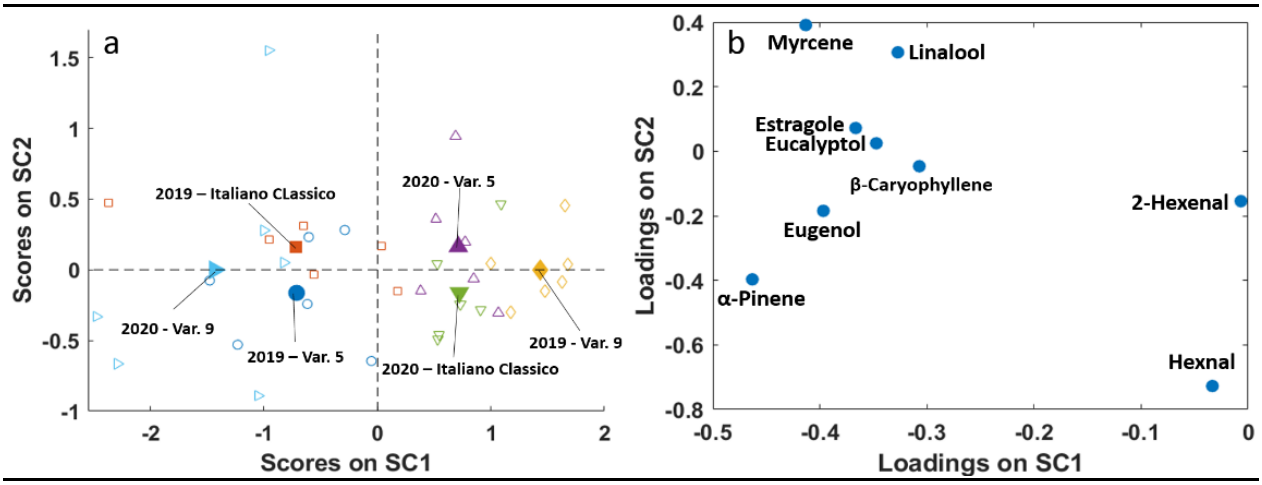


Figure 4-6. SCA of the effect matrix interaction "year x chemotype". (a) SC1 vs. SC2 scores plot with projected residuals (empty symbols); (b) variable loadings (SC1 vs. SC2).

The same pattern can be observed in Figure 4-7, which describes the effect of the interaction between the factors "cut" and "chemotype". In this case, the variation of factor "cut" is the one that strongly changes the chemical composition of samples characterized by the same chemotype, even if it does it to a lesser extent than the factor "year". High SC1 values correspond to a high 2-hexenal content, whereas low SC2 values are linked to high eugenol values.

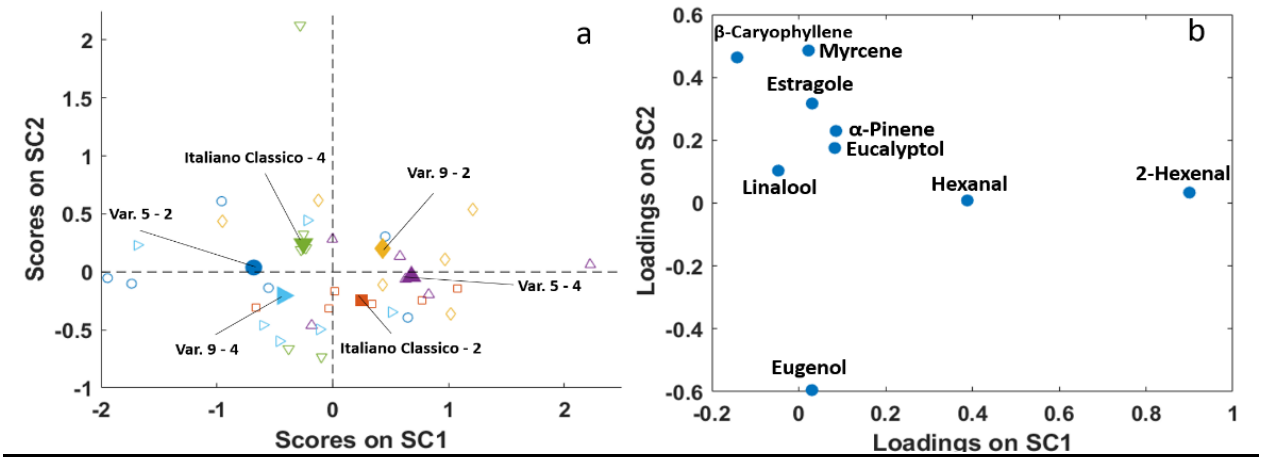


Figure 4-7. SCA of the effect matrix interaction "cut x chemotype". (a) SC1 vs. SC2 scores plot with projected residuals (empty symbols); (b) variable loadings (SC1 vs. SC2).

Considering the projected residuals, the differences are appreciable mainly in SC1, where Italiano Classico and Var. 9 show the same behaviour, being richer in floral/fruity flavours in cut 4 with respect to 2, while the opposite holds for Var. 5

4.1.2 Conclusion

The results obtained support the use of a fast-GC based electronic nose for rapid assessment of basil aroma; in fact, the main molecules perceived as persistent by olfactometry (GC/O) are identifiable and quantifiable. In agreement with previous literature, it has been observed that the aroma composition is not only a distinctive trait of chemotype, but the content of each specific molecule varies with agronomic year and cut period. On the one hand, this renders more problematic the choice of a specific chemotype to be cultivated to achieve a desired flavour profile; on the other hand, it may help focus on the chemotypes showing more stability with respect to the agronomic variability. In terms of percentage of variance, the cut affects the aroma less with respect to year and chemotype. The effect of year seems to be a bulk effect affecting the content more than the type of molecules found in the aroma.

4.2 Untargeted analysis of basil aroma

Here the study context and results are summarized, for more details, please refer to published paper number 2 in appendix 1.

The possibility to observe the complete chromatogram in an unsupervised way was the natural progression to fully benefit from the potential of the fast GC method. To this aim, the raw chromatographic signals, acquired in a very short time (110 s) were analysed together, after concatenation of the respective data matrices, according to a low-level data fusion approach [16, 17]. Furthermore, a higher number of basil samples collected from 2019 to 2021 (this year was not previously considered) were measured, while the number of chemotypes (chemotypes) studied was increased.

As pointed out, in this second study, the focus was on the extraction of reliable chemical information from the raw signals aided by proper data analysis and preprocessing tools. In this way, without the need and the effort of identifying and quantifying the specific markers, was nonetheless possible to study the different factors linked to production aspects and their influence on the product quality. This kind of approach could be easily and rapidly exported to other products where to acquire the knowledge of which individual molecules are present is more challenging or time consuming.

Multivariate data analysis pipeline included: proper preprocessing, exploratory analysis by Principal Component Analysis (PCA), and ANOVA Simultaneous Component Analysis (ASCA) [14] to assess the effect of chemotypes, cuts period and harvesting years (2019, 2020 and 2021) on basil aroma.

4.2.1 Results and Discussion

4.2.1.1 PCA Exploratory Analysis

In this first exploratory analysis, the aim was to obtain a general overview of the variation of the basil aroma. Punctual considerations of the influence of harvested year, chemotype and cut could not be conducted, since it was not possible to plain a systematic sampling beforehand, due to company and producer constrains. Three principal components were considered according to their explained variances (58%). In Figure 4-8, the PC1 vs. PC2 score plot is reported, representing the different basil samples with different symbols and colour as function of harvesting year and basil chemotype (Figure 4-8a) or cut and basil chemotype (Figure 4-8b).

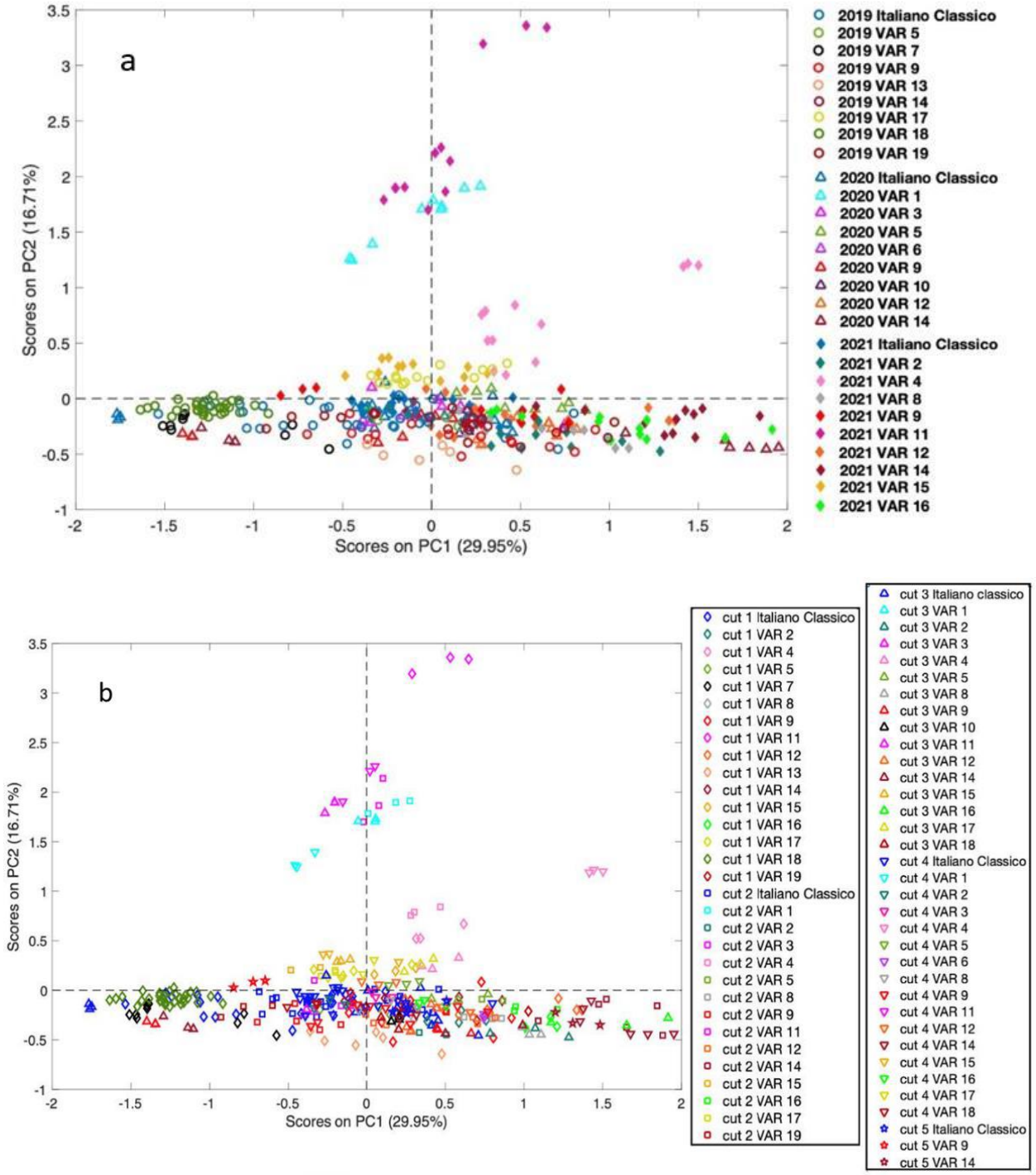


Figure 4-8. PC1 vs PC2 score plots of basil samples. (a) Different symbols were used for each harvesting year (2019: circles; 2020: squares; 2021: triangles) and distinct colours for each basil chemotype. (b) Different symbols were used for each cut (first: diamonds; second: squares; third and fourth: upwards and downwards triangles, respectively; fifth: stars) and distinct colours for each basil chemotype.

From the score plot of the first two components, it is difficult to highlight a clear separation of samples according to chemotypes, due to the slight differences in the flavour pattern among commercial chemotypes that belong to the same species (*O. basilicum*). However, interesting information can be pointed out. In particular, the VAR 1 (harvested only in 2019) and VAR 11 (harvested only in 2021) samples have the highest PC2 score values and leads to their separation from the other samples (Figure 4-8a). These chemotypes also present a trend, from higher to lower score values, according to their different cut (Figure 4-8b). Another peculiar chemotype

seems to be VAR 4 (harvested only in 2021), with positive scores for both PC1 and PC2. This chemotype shows differences in aroma according to different basil cuts as well.

As far as the other samples are concerned, they are distributed along the first principal component, which seems to be the most responsible for the differences in the separation between the VAR 14 samples (higher positive PC1 score values) and first cut of VAR 7, VAR 18 and Italiano Classico (negative PC1 score values).

Furthermore, the in-depth analysis of the figure shows that two samples belonging to the third cut of VAR 16 (higher PC1 score values) seem to have quite a similar aroma profile to VAR 14.

No further observations to assess any pattern can be performed considering the different basil cuts, years, and chemotypes, since it is not certain what the real cause is as some chemotypes were measured only in one year. The score plot of the third component (Figure 4-10) highlights the differences among the first basil cut of the VAR 8 and VAR 17 samples (higher positive score values) with respect to all the others.

From the PC1 loading plot (Figure 4-9a), for both MXT5 and MXT17 columns, it is possible to point out that, with almost all the loadings values being positive (from 40 to 110 s), the separation between the VAR 14 samples and the other basil chemotypes is mainly due to a global higher concentration of aroma compounds in these samples, and roughly speaking, most of the samples harvested in 2021 (positive PC1 score values) seem to present a similar trend.

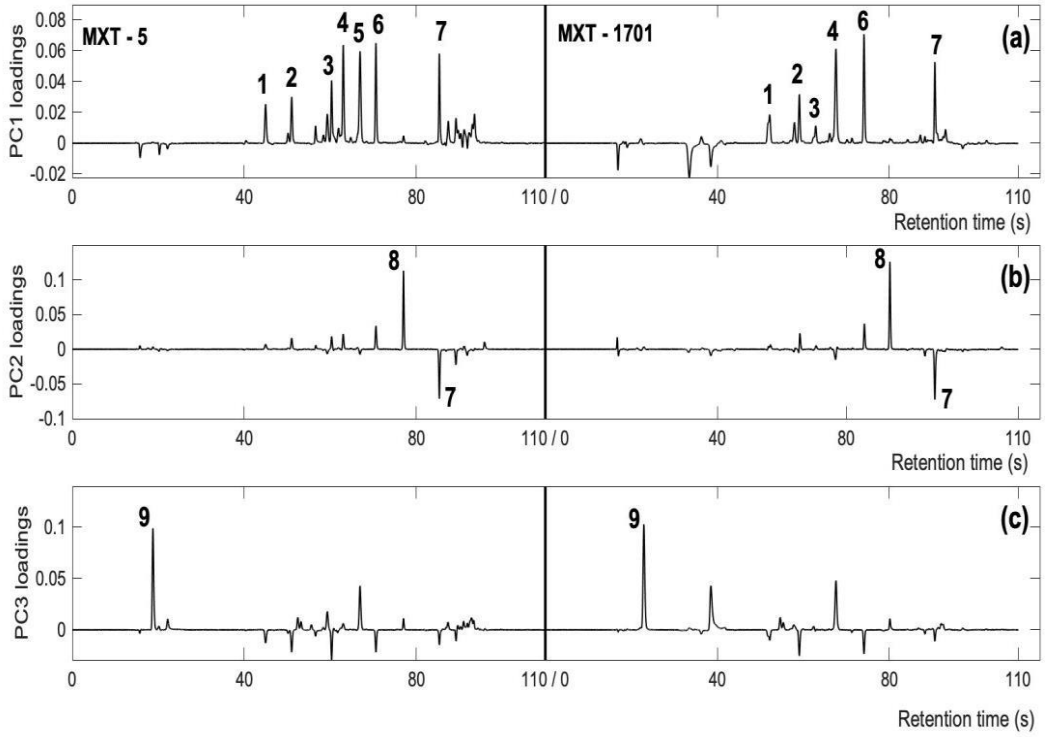


Figure 4-9. PC1, (b) PC2 and (c) PC3 loading plots. Numbered peaks correspond to the volatile compounds putatively identified on the basis of Kovats's relative retention indices: (1) hexanal, (2) 2-hexanal, (3) 5-methylfurfural, (4) myrcene, (5) eucalyptol, (6) linalool, (7) β-caryophyllene, and (8) eugenol (9) not identified.

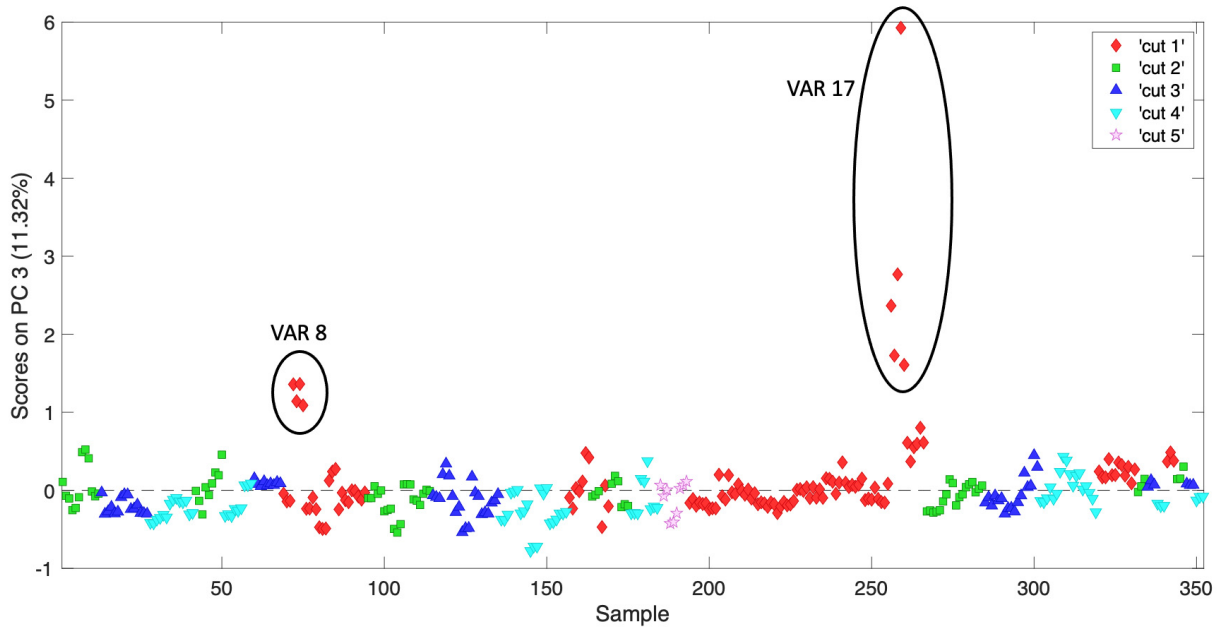


Figure 4-10. PC3 scores vs number of samples.

Notwithstanding the aim of the present study, which is to make a fast model to discriminate basil samples with an untargeted approach, some considerations on the presence of some chemical compounds can be presented based on our previous study. Regarding the second principal component Figure 4-9b), which is mainly responsible for the separation of VAR 1 and VAR 11 from the others, the same chromatographic regions (Rt, retention time: 76.8 s and 85.3 s for MXT-5 and 79.9 s and 90.4 s for MXT-17), for both the MXT-5 and MXT-17 columns, with the same trend (loadings value and sign), are relevant. Thus, both the estragole (Rt: 76.8 s and 79.9 s in MXT-5 and MXT-1701, respectively) and eugenol compounds (Rt: 85.3 s and 90.4 s in MXT-5 and MXT-1701, respectively), with high positive and negative loading values, respectively, are important to characterize VAR 1 and VAR 11. However, the samples belonging to these two chemotypes, presented a particular aroma, probably due to the presence of anethole, which co-elutes with estragole in both column separations.

As regards the third principal component (Figure 4-9c), unassigned compounds (in the first 40 s of both columns), which have positive loadings, seem more abundant in the VAR 8 and VAR 17 samples (located at positive scores values). Hence, further investigation will be conducted for the identification of these volatile compounds.

Notwithstanding the overall interpretation of PCA results, which offered some insights, more specific information is difficult to gain, since the contributions to variance of all the investigated factors (i.e., year, chemotype and cut) overlap. Therefore, after this preliminary investigation, the ASCA methodology was used to systematically assess the influence of each factor and their interaction on the basil aroma profile.

4.2.1.2 ASCA results

The first ASCA model was computed according to the regular experimental design that could be obtained limiting the analysis to only three chemotypes. The original data matrix variation was split in eight submatrices: three corresponding to the main effect of each experimental factor, three accounting for the effect of each second-order interaction, one describing the effect of the third-order interaction and one holding the residuals. The significance of all these effects was assessed by performing a permutation test, whose results are shown in Table 4-2. The p-value of all the inspected factors and interactions was lower than 0.001. However, the factors “chemotype” and “year” explained most of the data variance (39.9% and 24.8%, respectively), suggesting their higher influence on the aromatic profile of basil compared to the factor “cut”. This can also be observed by the fact that explained variance values of interactions including “cut” are

systematically lower than values related to interactions in which “cut” is not involved. Additionally, the third-order interaction effect explains less than 3% variance.

Table 4-2 - Explained variance and p-values for main factors and their second and third order interactions.

Factor	Explained Variance (%)	*p*
Chemotype	39.9	<0,001
Year	24.8	<0,001
Year x Chemotype	8.5	<0,001
Year x Cut	7.2	<0,001
Cut	2.9	<0,001
Chemotype x Cut	2.5	<0,001
Year x Chemotype x cut	2.8	<0,001

Afterwards, the ASCA algorithm performed a SCA on each effect matrix individually, with the aim of interpreting the observed variation.

Figure 4-11a shows the score plot for the factor “year”. The first component (SC1), which explains 67.7% of the total variance, describes the difference between the samples harvested in 2019 and the samples harvested in 2020 and 2021. The loadings plot of the first component, shown in Figure 4-11b, explains this difference. In fact, the 2020 and 2021 samples have a richer aroma profile, as the concentration of the compounds between 40 and 110 s, associated with statistically significant loadings, are higher compared to 2019 samples. On the other hand, 2019 samples present higher concentrations of unassigned peaks before 40 s highlighted by the MXT-1701 column, confirming the need of further investigation for their identification.

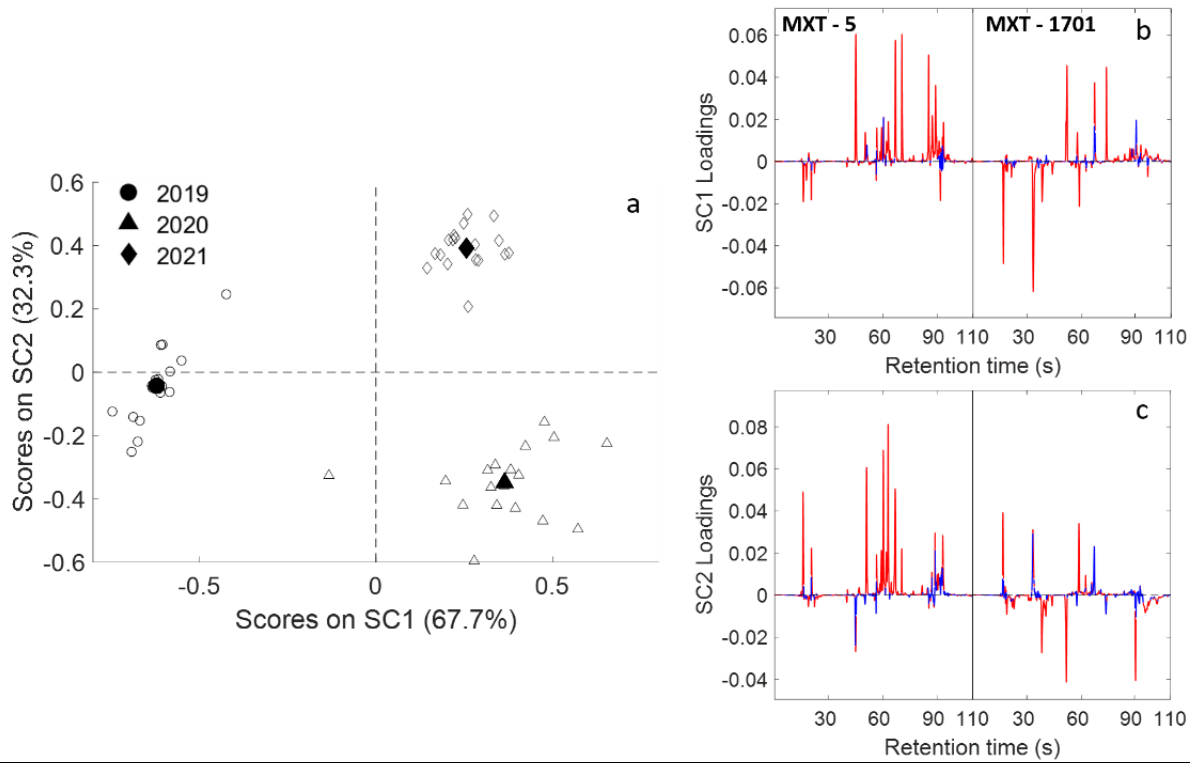


Figure 4-11. SCA for the effect of the factor “year”. (a) SC1 vs. SC2 score plot. Empty symbols represent the projected residuals; (b) SC1 and (c) SC2 loadings plot. In the loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated with loadings statistically indistinguishable from zero.

The second component (SC2) and the related loadings plot (Figure 4-11c) show how the 2021 samples (positive scores values) present lower peaks in MXT-1701 that can be ascribed to 2-hexanal and β-caryophyllene (negative loadings values), but higher peaks assigned to all other compounds.

Figure 4-12a shows the score plot for the factor “chemotype”. It can be observed that most of the explained variance (96.3%) describes how VAR 14 is different compared to Italiano Classico and VAR 9. Indeed, as shown by the loadings plot in Figure 4-12b, VAR 14 presents higher concentrations of all the chromatographic peaks, suggesting a richer aroma profile with respect to the other two chemotypes. SC2, even though the related explained variance is extremely low (3.7%), mainly shows how VAR 9 has more β-caryophyllene than Italiano Classico (Figure 4-12c), as their peaks are basically the only ones that had statistically significant results.

The results of the SCA for the effect of the interaction “year x chemotype” were reported in Figure 4-13. In the score plot (Figure 4-13a), it can be observed that SC1 describes the difference among VAR 14 samples throughout the years. In detail, the VAR 14 samples collected in 2020 presented a higher concentration of all aroma compounds compared to the ones collected in 2019 and 2021, as assumed by the loadings plot shown in Figure 4-13b. As regards Italiano Classico, the best year in terms of intensity of aroma profile is 2019, whereas for VAR 9, the years 2019 and 2021 were better than 2020.

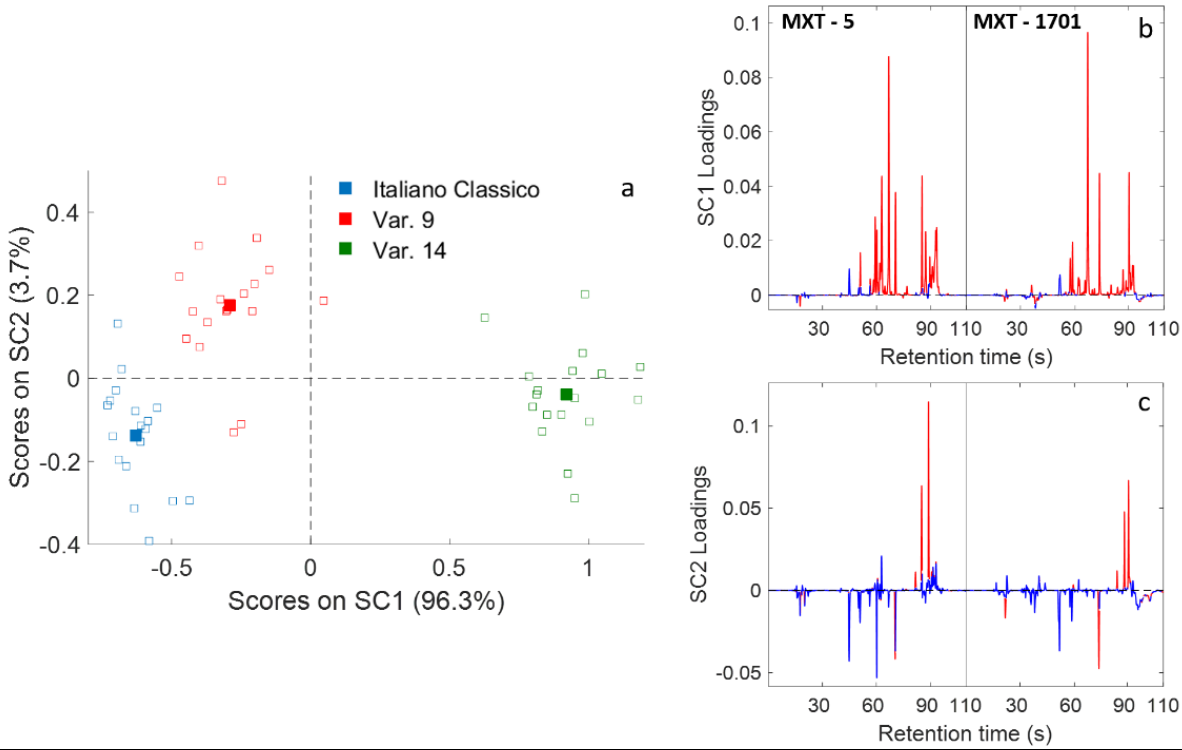


Figure 4-12. SCA for the effect of the factor “chemotype”. (**a**) SC1 vs. SC2 score plot. Empty symbols represent the projected residuals; (**b**) SC1 and (**c**) SC2 loadings plot. In the loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated with loadings statistically indistinguishable from zero.

It can also be observed how VAR 14 appears to change more over time, having a higher variation through the years than the other two chemotypes.

Moreover, Italiano Classico is the basil chemotype that presents the lowest variability among its replicates. In fact, red and green samples in the score plot (VAR 9 and VAR 14, respectively) are more spread and farther apart, especially along SC2. This limits further comments about the difference between the years 2020 and 2021 with respect to the Italiano Classico samples (blue triangles and diamonds in Figure 4-13a, respectively), which is due to the statistically significant peaks between 50 and 70 s, linked to most of the aromatic compounds.

Regarding the factor “cut”, the SCA showed how samples collected during cut 2 detain a richer aroma profile than samples acquired during cut 4. However, according to the authors, since this factor explained less than 3% of the total variance, these results are not relevant compared to the ones described above. Both for this reason and for the sake of brevity, plots related to the factor “cut” were not shown.

The second ASCA model was computed considering only samples collected in 2021. In this case, it was possible to build a balanced design, including nine chemotypes and three cuts (see

paper 1) The data matrix was partitioned in four submatrices: two corresponding to the main effect of each experimental factor, one describing the effect of the second-order interactions and the residuals matrix. The results of the permutation test for the significance of the effects are shown in Table 4-3. As for the first ASCA model, also in this case, all the factors and their interactions were significant ($p < 0.001$). Furthermore, the explained variance for the factor "cut" (6.9%) was significantly lower than the variance explained by the factor "chemotype" (63.5%), suggesting, once again, the small impact of plant age on the basil aroma profile.

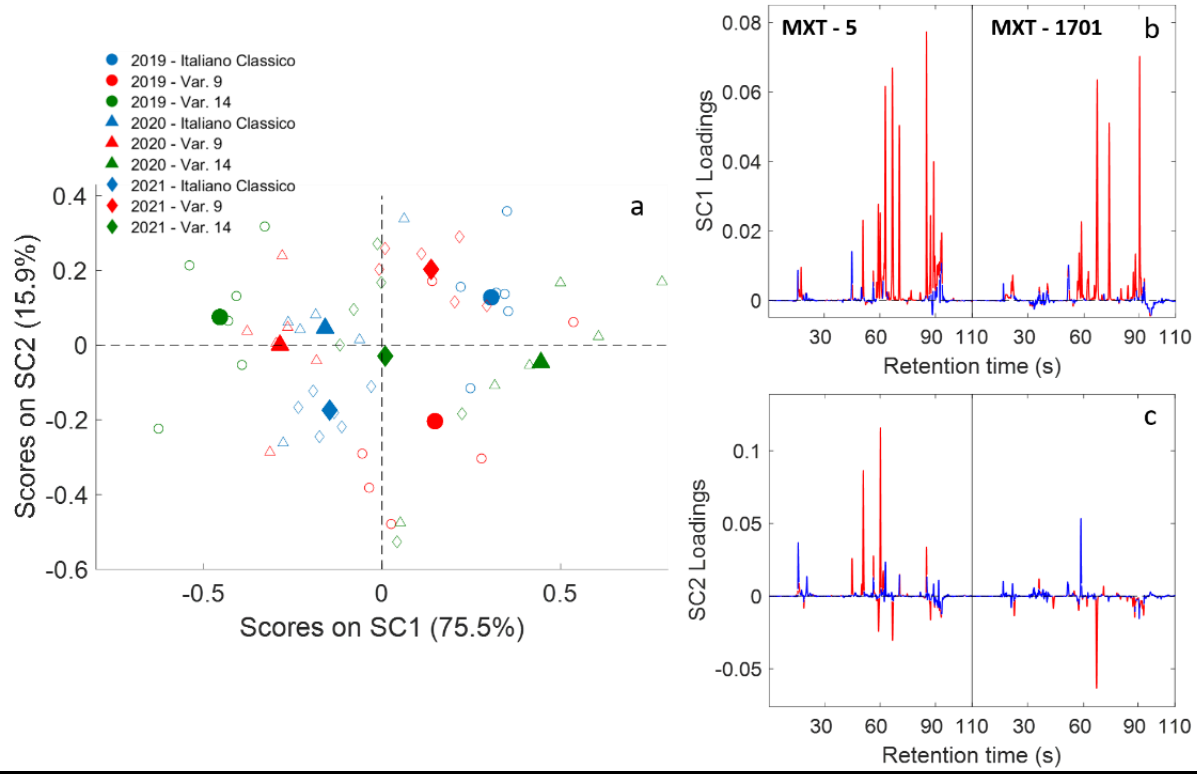


Figure 4-13. SCA for the effect of interaction "year x chemotype". (a) SC1 vs SC2 Score plot. Empty symbols represent the projected residuals; (b) SC1 and (c) SC2 loadings plot. In (a) distinct colours refer to different chemotypes (blue - Italiano classico; red - VAR 9; green - VAR 14), whereas different symbols refer to different harvesting years (circles - 2019; triangles - 2020; diamonds - 2021). In loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated to loadings statistically indistinguishable from zero.

Table 4-3. Explained variance and p-values for main factors and their second order interactions related to the ASCA model.

Factor	Explained Variance (%)	*p*
Chemotype	63.5	<0,001
Chemotype x Cut	20.3	<0,001
Cut	6.9	<0,001

The results related to SCA on the "chemotype" effect matrix are shown in Figure 4-14.

From the score plot (Figure 4-14a), it is clear how the first principal component shows the difference between VAR 4 and all the other chemotypes. In the loadings plot (Figure 4-14b), it is shown that the peak that is responsible for this difference can be ascribed to myrcene, of which VAR 4 is particularly rich. Observing SC2 scores and loadings (Figure 4-14c), it can be concluded that VAR 14 and VAR 16 present the richest aroma profiles, whereas Italiano Classico and VAR 15 have the poorest profiles.

Figure 4-15a shows the frequency histogram of the SC1 scores values for the distinct levels of the factor "cut". Eucalyptol and β-caryophyllene are less present in cut 4 samples, and in general, they are the compounds responsible for describing the difference between cut 4 samples and cut 1 and 2 samples, as shown in Figure 4-15b.

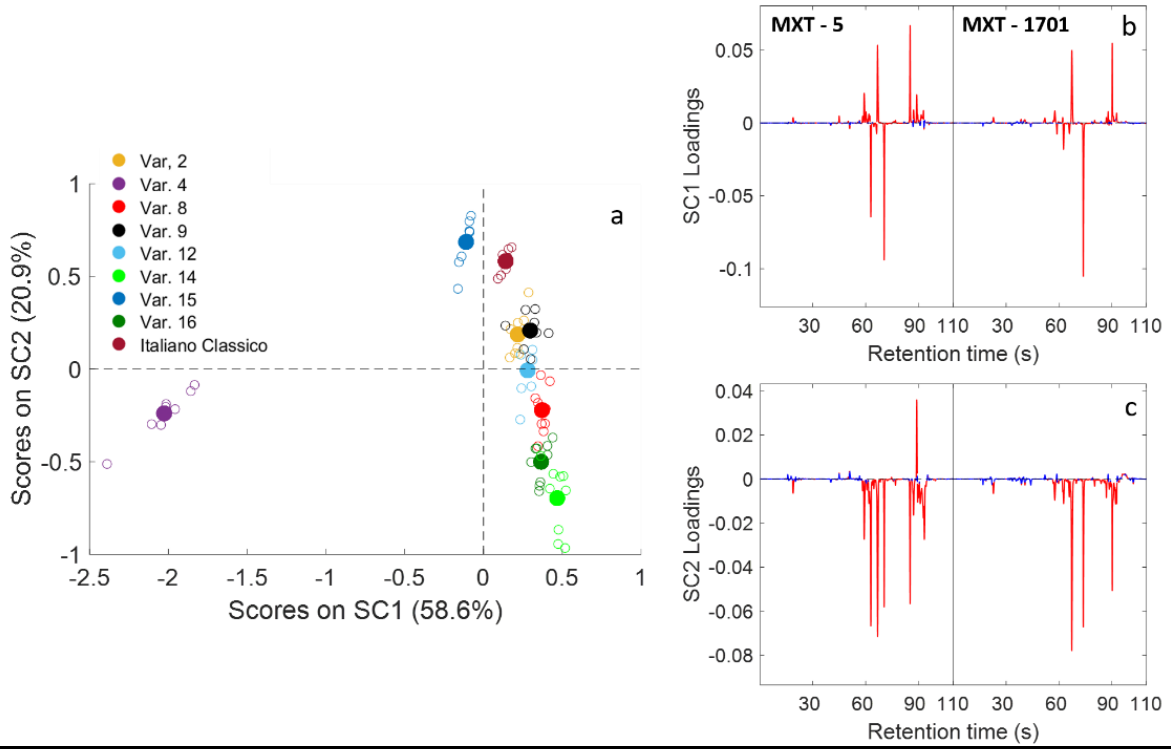


Figure 4-14. Results of ASCA performed on 2021 samples. SCA for the effect of factor "chemotype". (a) SC1 vs SC2 Score plot. Empty symbols represent the projected residuals; (b) SC1 and (c) SC2 Loadings plot. In loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated to loadings statistically indistinguishable from zero.

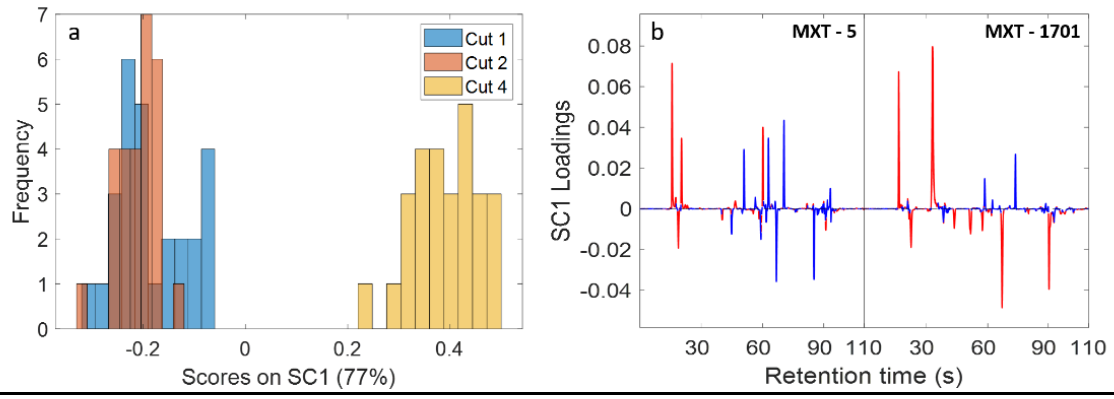


Figure 4-15. Results of ASCA performed on 2021 samples. SCA for the effect of factor "cut". (a) histograms of ASCA score frequency (with projected residuals) on SC1 for the distinct levels of factor "cut"; (b) SC1 Loadings plot. In loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated to loadings statistically indistinguishable from zero.

The ASCA results show how the entire aromatic profile has a significant influence in the discrimination of samples according to the investigated factors (i.e., years, chemotype and cut), highlighting the presence of new potential biomarkers (for instance the species with retention time in the first 30 s of the chromatogram or the ones falling in the area between the retention of 2-hexanal and 5-methylfurfural), which have not been quantified in this study, but that could be relevant in further investigations. For the sake of clarity, an example signal fingerprint with all the chemical analytes, putatively identified for both the chromatographic separations, is reported in Figure 4-16.

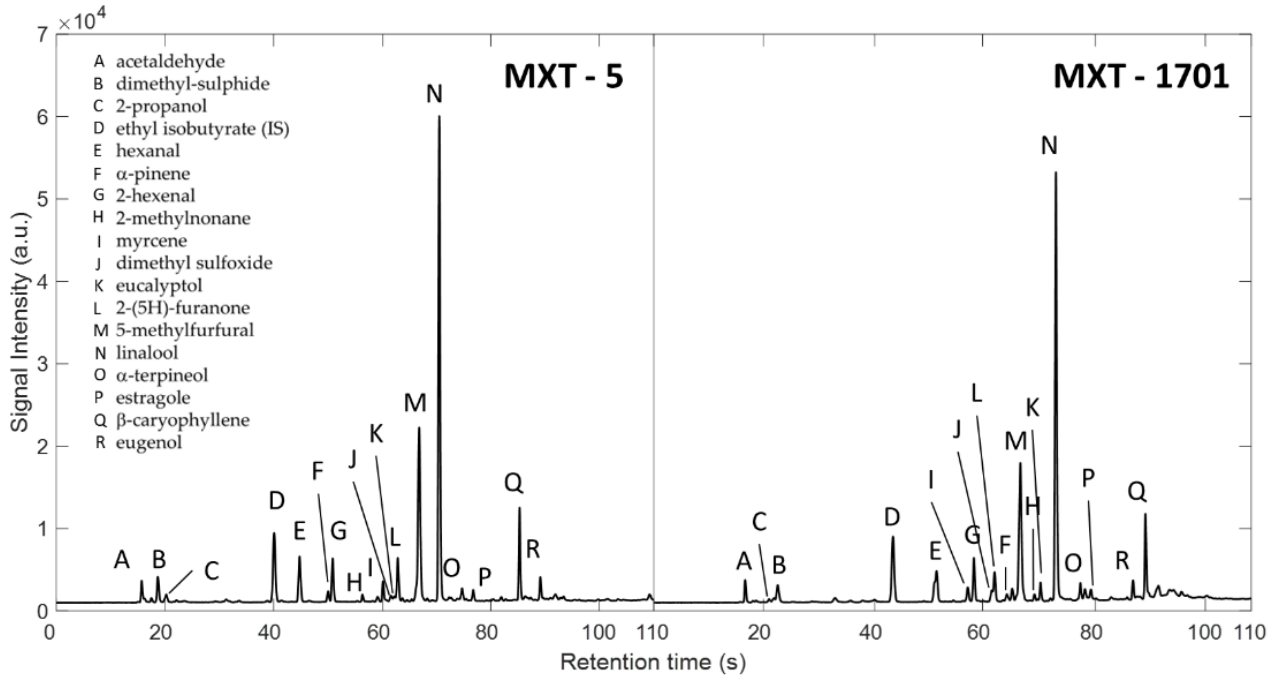


Figure 4-16. Chromatograms of Italiano Classico chemotype obtained by elution on columns MXT-5 and MXT-1701 of Heracles II.

4.2.2 Conclusions

In this second study, the development of a fast analytical screening strategy based on ultra-fast chromatography e-nose and multivariate analysis was proposed as a useful tool for quality control of food. The proposed approach, relying on the simultaneous analysis of the chromatographic profiles coming from two GC-columns of different polarity, permits to fully explore the volatile profile of foodstuff and may represent a fast and simpler alternative to other chromatographic techniques. The chemical identification and quantification of the single chemical species, responsible for differentiation of the studied food products, can be undertaken on a few samples at a second time if necessary. In fact, once the main chromatographic peaks, most responsible for the differentiation between samples, have been underlined, their respective chemical species can be identified with a considerable reduction in costs and analysis time.

This approach was applied on the analysis of basil samples involved in the production of Italian pesto sauce, where the whole GC-FID e-nose signals, coming from two columns with different polarity, were fused and used as a fingerprint of the aroma profile. The obtained results highlighted the possibility to differentiate basil samples based on the three investigated factors, years, cut and chemotype, taking also into account the interactions among them. The low-level data fusion approach allowed computing a single ASCA model, which effectively pointed out the different significant peaks between the two columns considered, thus underlining that enhanced information may be gained.

The knowledge of the influence of the investigated factors on the quality of basil is very important, since it may allow a company to achieve useful information both to plan future campaign strategies for the acquisition of the raw materials and to improve the quality of the final pesto sauce.

1 Eileen MK, Emily DN. Variations in phenolic composition and antioxidant properties among 15 basil (Ocimum basilicum L.) cultivars. Food Chem. 2011; 128: 1044–1050.

2 Southwell IA, Russel MF, Davies NW. Detecting traces of methyl eugenol in essential oils: Tea tree oil, a case study. Flavour Fragr. J. 2011; 26: 336–340.

3 Murarikova A, Tazky A, Neugebaureova J, Plankova A, Jampilek J, Mucaji P et al. Characterization of Essential Oil Composition in Different Basil Species and Pot Cultures by a GC-MS Method. Molecules. 2017; 22: 1221.

4 Lee SJ, Umano K, Shibamoto T, Lee KG. Identification of volatile components in basil (Ocimum basilicum L.) and thyme leaves (Thymus vulgaris L.) and their antioxidant properties. Food Chem. 2005; 91: 131–137.

5 Leonardos G, Kendall D, Barnard N, Odor Threshold Determinations of 53 Odorant Chemicals. J. Air Pollut. Control. Assoc. 1969; 19: 91–95.

6 Plotto A, Margaria CA, Goodner KL, Baldwin EA. Odour and flavour threshold for key aroma components in an orange juice matrix: Terpenes and aldehydes. Flavour Fragr. J. 2004; 19: 491–498.

7 Salvadeo P, Boggia R, Evangelisti F, Zunin P. Analysis of the volatile fraction of "Pesto Genovese" by headspace sorptive extraction (HSSE). Food Chem. 2007; 105: 1228–1235.

8 Bertoli A, Lucchesini M, Mensuali-Sodi A, Leonardi M, Doveri S, Magnabosco A, Pistelli L. Aroma characterization and UV elicitation of purple basil from different plant tissue cultures Food Chem. 2013; 141: 776–787.

9 Zlotek U, Mikulska S, Nagajek M, Swieca M. The effect of different solvents and number of extraction steps on the polyphenol content and antioxidant capacity of basil leaves (Ocimum basilicum L.) extracts. Saudi J. Biol. Sci. 2016; 23: 628–633.

10 Lu Y, Gao B, Chen P, Charles D, Yu L. Characterisation of organic and conventional sweet basil leaves using chromatographic and flow-injection mass spectrometric (FIMS) fingerprints combined with principal component analysis. Food Chem. 2014; 154: 262–268.

11 Jordán MJ, Quílez M, Luna MC, Bekhradi F, Sotomayor JA, Sánchez-Gómez P et al. Influence of water stress and storage time on preservation of the fresh volatile profile of three basil genotypes. Food Chem. 2016; 13: 169–177.

12 Fratianni F, Cefola M, Pace B, Cozzolino R, De Giulio B, Cozzolino A et al. Changes in visual quality, physiological and biochemical parameters assessed during the postharvest storage at chilling or non-chilling temperatures of three sweet basil (Ocimum basilicum L.) cultivars. Food Chem. 2017; 229: 752–760.

13 Smilde AK, Jansen JJ, Hoefsloot HC, Lamers RJA, Van Der Greef J, Timmerman ME. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. Bioinformatics 2005; 21: 3043–3048.

14 Jansen JJ, Hoefsloot HC, van der Greef J, Timmerman ME, Westerhuis JA, Smilde AK. ASCA: Analysis of multivariate data obtained from an experimental design. J. Chemom. J. Chemom. Soc. 2005; 19: 469–481.

15 Acree TE. GC/olfactometry GC with a sense of smell. Anal. Chem. 1997; 69: 170A–175A.

16 Giannoukos K, Giannoukos S, Lagogianni C, Tsitsigiannis DI, Taylor S. Analysis of volatile emissions from grape berries infected with Aspergillus carbonarius using hyphenated and portable mass spectrometry. Sci. Rep. 2020; 10: 1–11.

17 *Torres MN, Valdes NB, Almirall JR. Comparison of portable and benchtop GC–MS coupled to capillary microextraction of volatiles (CMV) for the extraction and analysis of ignitable liquid residues. Forensic Chem. 2020; 19: 100240.*

5EXPLOITING PESTO SAUCE: A DATA FUSION APPROACH

In R&D an important aspect is to assess which analytical technique to adopt in routine analysis reaching the best compromise among costs, time and personnel expertise required. The main aim is to dispose of fast and easy to operate methods to afford a larger number of samples to be routinely analysed.

In this study, three analytical techniques were considered: GC-FID e-nose (successfully applied to inspect basil aroma), head space gas chromatography ion mobility spectroscopy (HS-GC-IMS) and near infrared spectroscopy (NIRS).

HS-GC-IMS [1] is very sensitive but requires complex and time-consuming data elaboration routines. The GC-FID e-nose [2] is enough fast but requires trained people and a laboratory context. The NIRS [3] is a very rapid and easy-to-use technique with high potential to application in an industrial context. In fact, it allows quick evaluation of the product characteristics also in-situ/on-line, but its capability to "see" different aromas needs to be verified.

The objective here was to evaluate the capacity of each technique to differentiate the classes of pesto, accordingly the data analysis pipeline included exploratory analysis, and applying discriminant analysis (also coupled to variable selection) on each data set. Then, it was also evaluated the discriminant capacity of all the techniques used together with a low-level multi-block data fusion approach.

The obtained results indicates that GC-FID e-nose was more efficient in separating the pesto classes, followed by NIRS that was shown to be promising in differentiating the pesto categories. A variable selection applied to each single analytical technique helped to interpret the causes of the differences between pesto samples.

The combined data from GC-FID e-nose, the HS-GC-IMS and NIRS did not give a significant increase in discrimination performance, also after the variable selection application. However, it gave useful information to understand which analytical technique could be useful in pesto characterisation.

An important take home message was the confirmation of the ability of the tested methods that measure the aroma profile, to characterize pesto classes but, more interestingly, that also NIRS can successfully be used to distinguish pesto classes, with potential future applications in industrial environment.

5.1 Materials and methods

5.1.1 Sampling

Twenty-six samples of "*Pesto alla Genovese*" produced by Barilla in Rubbiano plant facility, were selected from the whole 2021-year production period and analysed. They covered the use of three different basil categories. For confidentiality reasons the three basil categories will be just reported as class1, class2 and class3.

The analytical methods details have been reported respectively: for GC-FID e-nose in paragraph 4.5.2, for HS-GC-IMS in paragraph 4.6 and for NIRS in paragraph 4.7.

5.2 Results and discussion

Results will be presented first per single techniques also describing the data preprocessing and issues posed by each type of data, and afterwards the data fusion results will be presented.

5.2.1 HS-GC-IMS data

The raw data from HS-GC-IMS, are *per* single sample a landscape reporting signal intensity as function of the two distinct separation dimensions, retention time (chromatographic dimension) and drift time (ion mobility dimension). The data size is huge, and several issues must be faced such as shift in drift time dimension.

5.2.1.1 Preprocessing steps

In Ion Mobility Spectrometry [4], that works at ambient pressure, the ionization of the analytes molecules passes through the ionization of water molecules naturally present in the ionization chamber. The water ions, $(H2O)^n(H3O)^+$, then exchange charge with the analyte molecules coming from the gas chromatographic column. So typically, it is visible a peak of the charged water (called Reactant Ion Peak, RIP) in the first part of the chromatogram, that decreases in intensity depending on the given charge. Because it is not useful in the elaboration, the RIP zone has been cut and removed from the chromatograms.

For computational reasons it was necessary to preliminary reduce the dataset size before further elaboration. The first step has been to retain just the informative part of the mass direction and remove empty regions: the final retained range in the drift time was from 8.5 to 18 ms. In the chromatographic direction the full period window was maintained, but the number of points was reduced collecting one point every ten. This was possible because the sampling frequency was enough high and the chromatographical profile was not altered after the downsizing. In this way the sample landscape dimensions were reduced from the original 6285 (Rt) x4500 (ms) to 629 (Rt) x1427 (ms) (Figure 5-1).

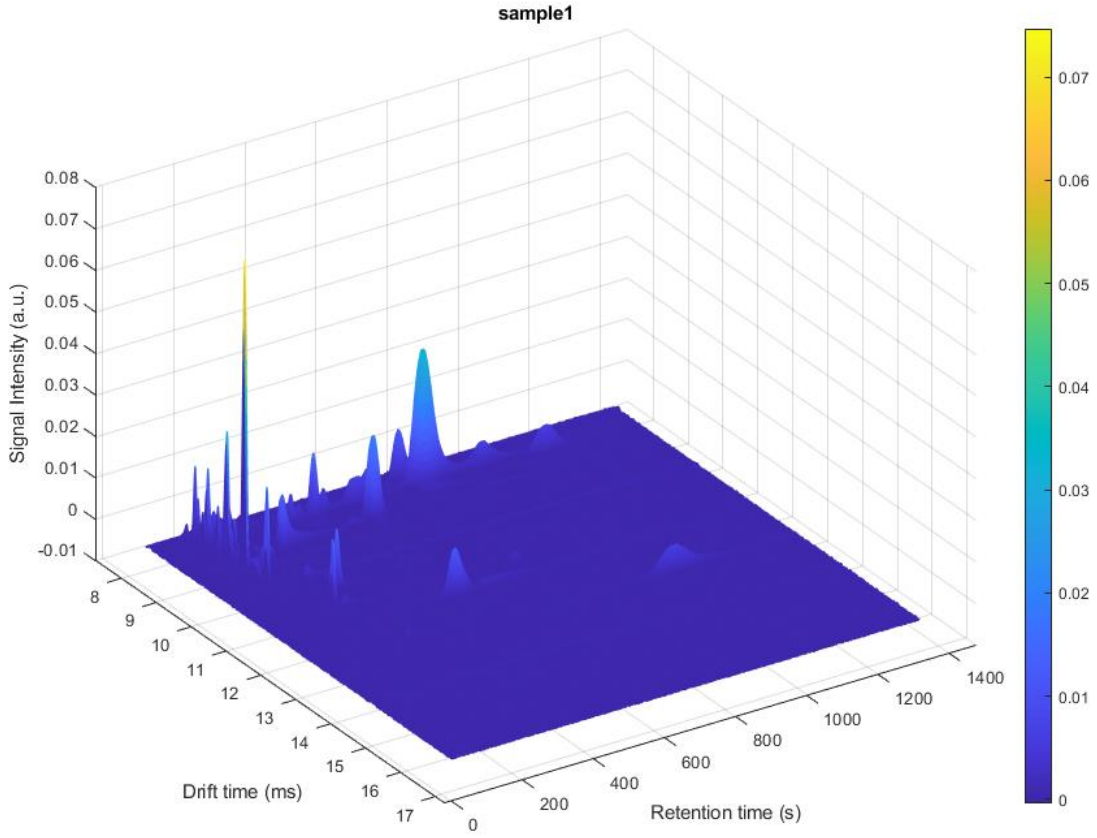


Figure 5-1 GC-IMS landscape for sample 1, after data reduction. The x-y axes are the two separative dimensions, chromatographical retention time and drift time, while the z axis reports the signal intensity.

After data reduction preprocessing was applied, this was decided based on inspection of the raw data as shown in Figure 5-2.

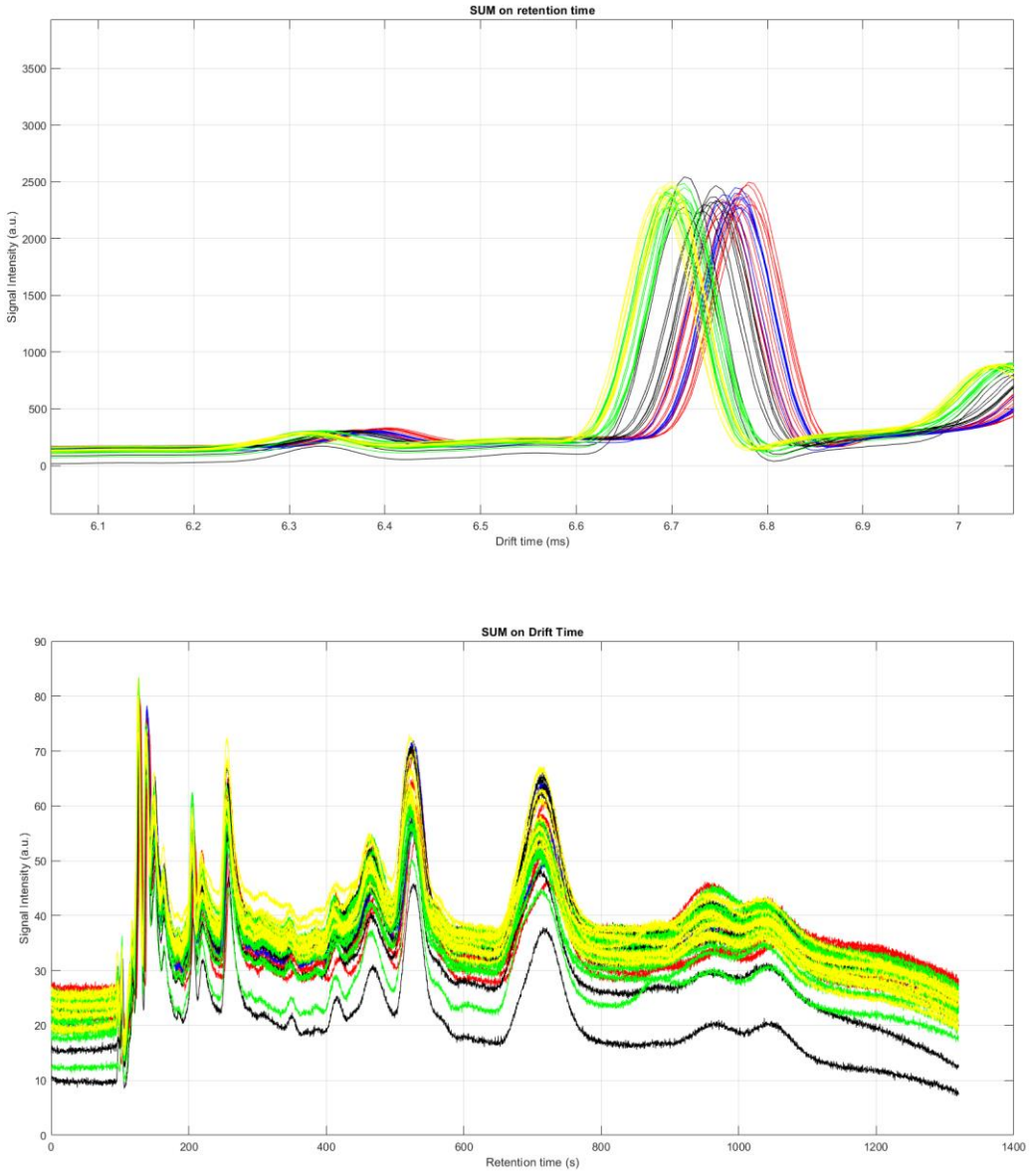


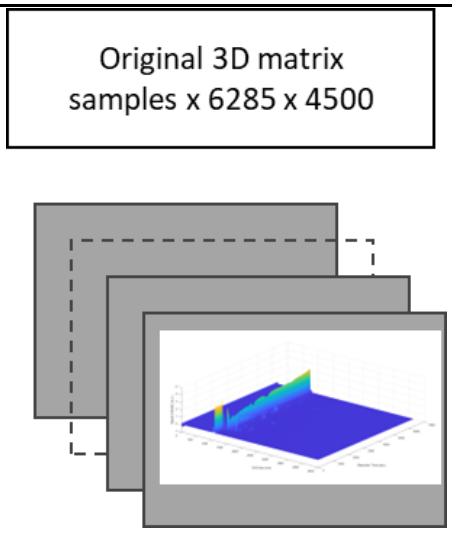
Figure 5-2. Top) drift time profile, obtained by taking the sum at each single drift time along the Rt dimension; day to day shift observed on RIP in mass (drift time) direction. Bottom) chromatographic profile, obtained by taking the sum at each single retention time along the drift time dimension; no shift is observable, but signals are noisy, in chromatographic direction.

In particular, day to day shift is present in mass direction (drift time) while noise and baseline are observable in chromatographic direction (retention time). In addition, normalization is required because absolute intensity is run dependent so to compare the different samples is better to switch to relative intensity profiles.

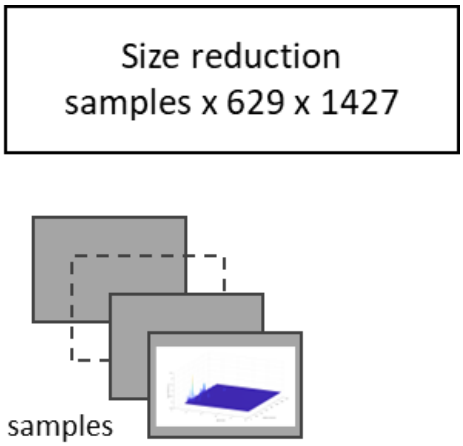
Misalignment in chromatographic direction seems not to be present, anyhow since MCR-ALS will be applied on the data unfolded along retention time dimension (in multiset modality, i.e. to each sample will correspond its own resolved chromatographic profiles one for each resolved component) shift in this dimension is of no concern.

Thus, samples were first aligned on drift time direction to compensate small shifts due to fluctuation in the ambient pressure between days. In fact, the ions mobility into the drift tube depends on the ambient pressure [5]. Alignment was done using the icoshift algorithm (see paragraph 3.2.1). Data were then smoothed with Savitzky-Golay filter, baseline corrected

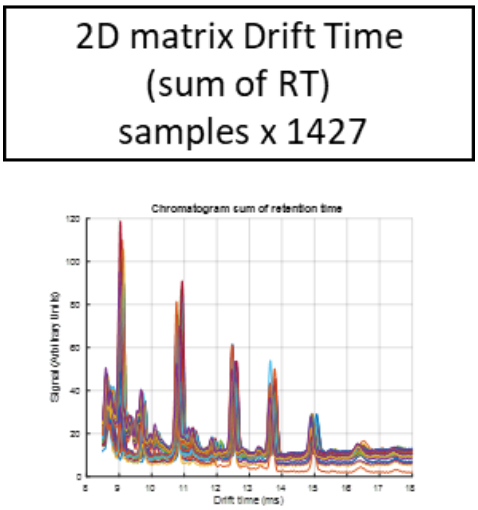
(Whittaker filter [6] 0.001, lambda 100) and then normalized (dividing by Euclidean norm). In Figure 5-3 the applied preprocessing steps are illustrated.



Size reduction in chromatographic direction and cuts in drift time direction



Sum on chromatographic direction to study misalignment on drift time direction



shifts are observable in drift time direction.

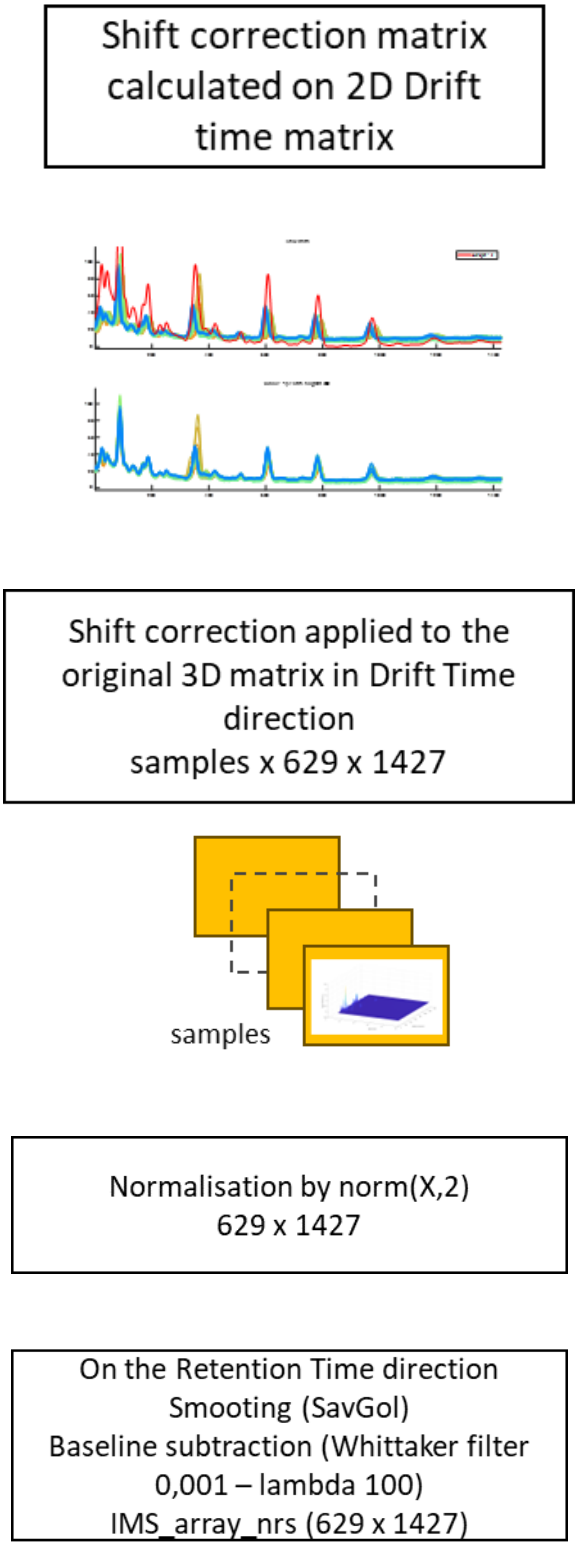


Figure 5-3. scheme of data processing for GC-IMS data.

5.2.1.2 Decomposition/Resolution by MCR

Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) (see chapter 3) was applied to the pre-processed HS-GC-IMS data. Typically, a peak peeking is done on the GC-IMS landscape for a representative sample to select manually the peaks present [7,8] then these are sought by the instrument software in all acquired samples and integrated. More recently, several

multivariate approaches were proposed in the literature [9,10,11]. Among them MCR-ALS [12,13] is appealing since it can resolve overlapped signals in “pure” components contributions considering the information of the second dimension, in our case the drift time related to ion mobility. In this way, it become possible to separate the single chemical components contribution present in the samples.

A preliminary MCR application on the whole multiset, i.e. all 26 samples (data not reported), showed sub-optimal, thus, to improve the separation performances, the chromatograms were divided in six intervals (interval one was discarded because do not contain any peaks). This is quite common to do when MCR is applied to hyphenated chromatographic techniques as well as to GC-IMS [14,15,16]. Thus, five matrices were prepared one for each interval, unfolding for all the samples the GC-IMS landscape row-wise with the drift time in columns and the retention time of all the samples concatenated in rows. These matrices were then decomposed by MCR-ALS. Non-negativity was imposed as constrain on both **C** and **S** matrices during ALS iterations.

For each interval have been retained only the MCR components that showed a clear peak profile as reported in Table 5-1, discarding components ascribable to baseline contributions. In Figure 5-4 it is shown an example of components resolved in one of the intervals.

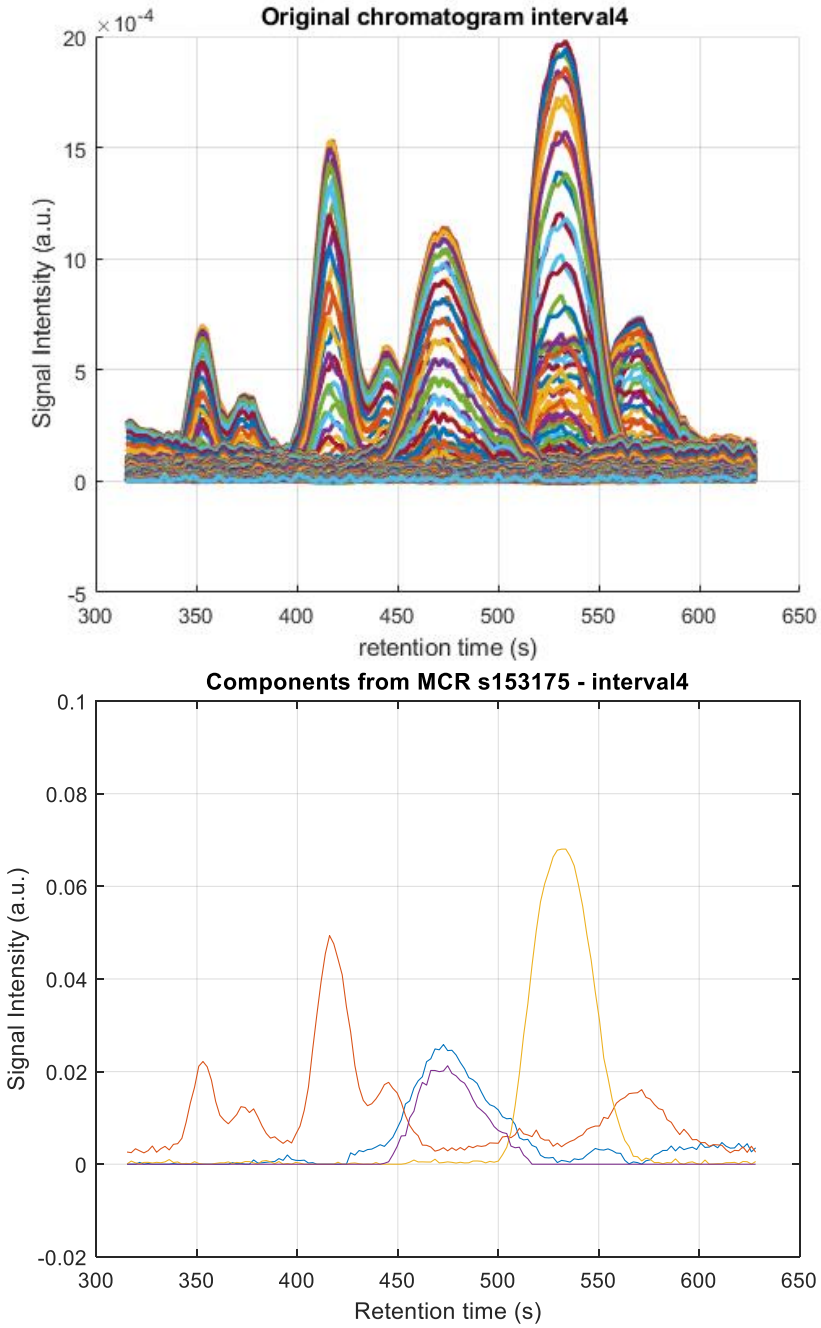


Figure 5-4. As an example, it is shown for one of the samples: in the upper figure are superimposed all the chromatograms, at the different drift times, for the retention time region corresponding to the interval 4; in the lower figure the chromatographic profiles of the resolved components selected in the same interval.

After the application of MCR-ALS routine, the peak areas of all the selected components were calculated and joined in a single data set, which was then input to further multivariate analysis.

Table 5-1. MCR components selected for each interval.

INTERVALS	Resolved components	Selected components for further data analysis
INT2	20	8
INT3	9	4
INT4	86	4
INT5	88	4
INT6	55	3

Firstly, an explorative PCA was done (data not reported) to have an overview on the sample's similarity/differences.

Then, data was split respectively in 31 calibration samples and 12 validation samples, and PLS-DA was applied (Figure 5-5).

To estimate the correct number of latent variables to be used in the PLS-DA models, cross-validation was performed with a venetian blind scheme using 10 splits. Six latent variables were selected.

The results are shown in Figure 5-5, where it can be observed that class 2 is well separated, while classes 1 and 3 are overlapped.

The confusion matrix in cross validation and in prediction are reported respectively in Table 5-2 and Table 5-3 Samples belonging to Class 2 are always correctly predicted, while the other classes have some misclassified especially none of the class 1 test samples is recognized as belonging to it. In Figure 5-6, are reported the predicted Y-value vs. N° of samples (test samples are separated by a vertical line) and the class threshold (horizontal red line). It is possible to observe that PLS-DA model can correctly allocate the test samples only for classes 2 and 3. Off course the low number of samples prevent any assessment of predictive performance, but as feasibility of the technique to reflect pesto type 2 and 3 results seem encouraging.

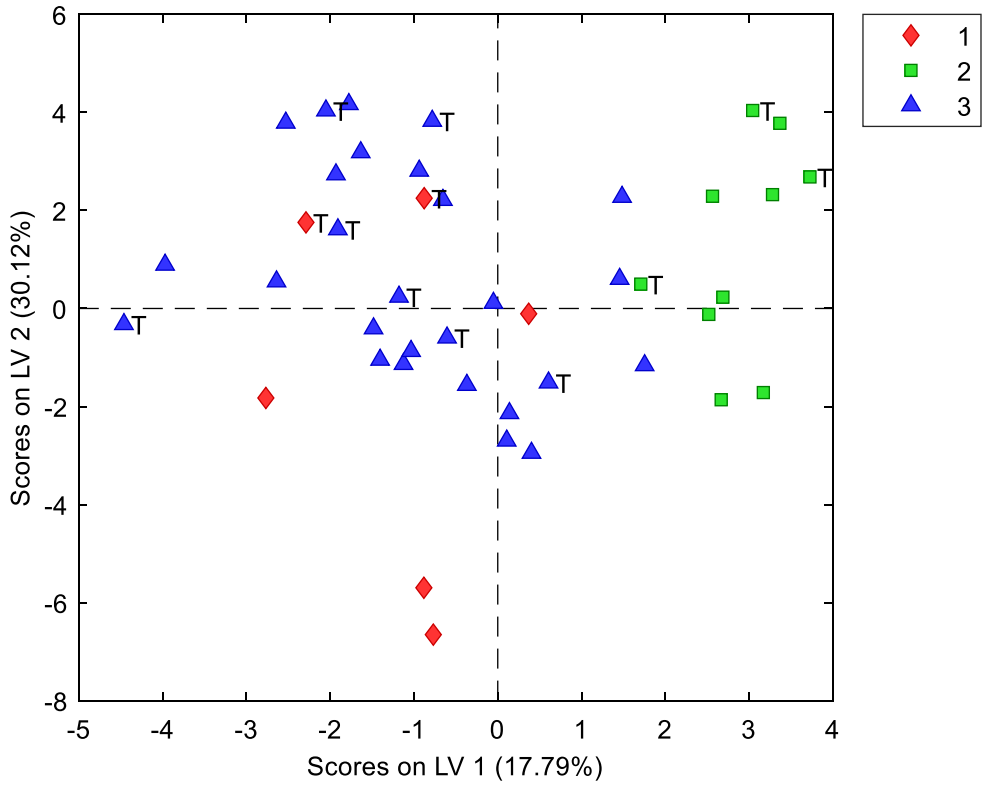


Figure 5-5. PLS-DA (model based on HS-GC-IMS) scores plot of all samples, coloured by classes; test set samples are indicated whit a T.

Table 5-2. Confusion matrix reporting the number of samples recognized in Cross Validation for each class (PLS-DA model obtained by HS-GC-IMS data)

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	2	0	5
Predicted as 2	0	7	0
Predicted as 3	2	0	5
unassigned	0	0	0

Table 5-3. Confusion matrix reporting the number of samples recognized in Prediction (test set) for each class (PLS-DA model obtained by HS-GC-IMS data)

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	0	0	0
Predicted as 2	0	3	0
Predicted as 3	2	0	7
unassigned	0	0	0

The VIP scores are reported in Figure 5-7 and indicate that to class 2 separation contribute most of the resolved components, indicating how this samples have quite different compositional profile respect to the other two classes. The components important for prediction of Class1 and 3 memberships are almost the same and this could explain the lower capability of the model to discriminate class 1 from 3.

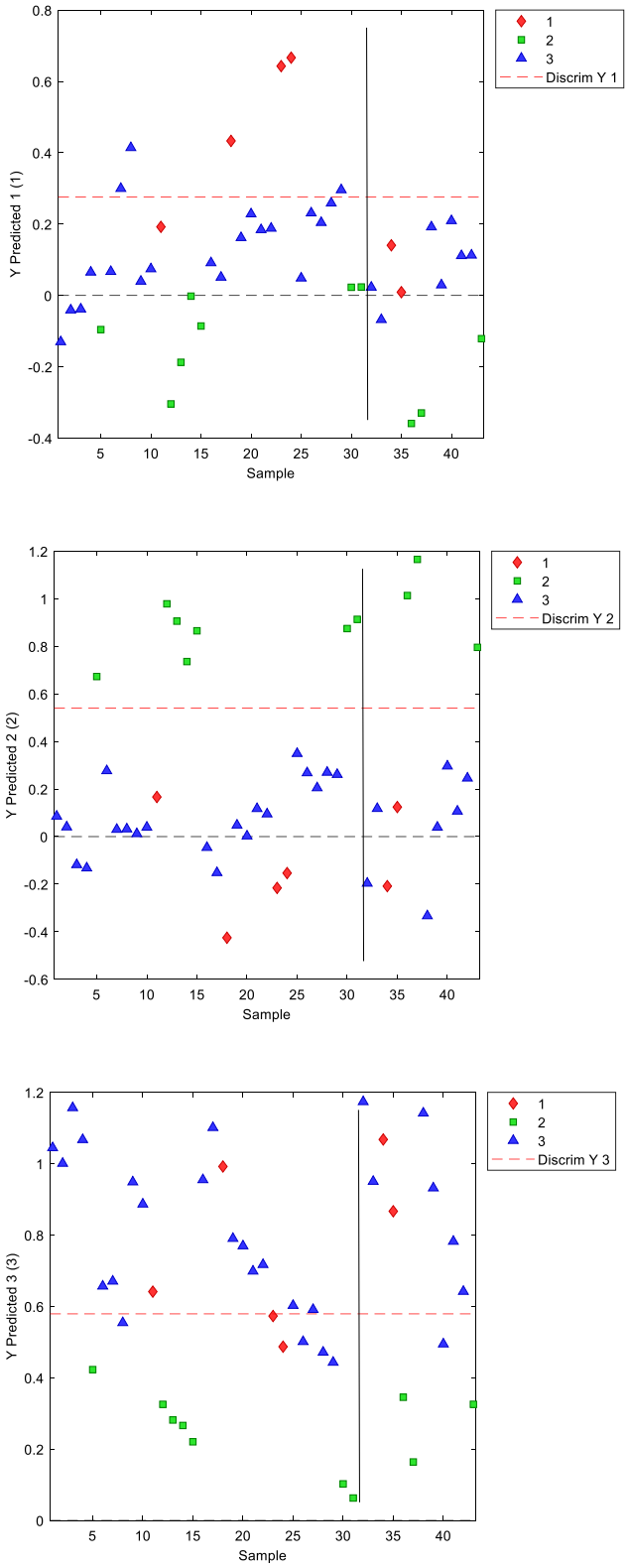


Figure 5-6. PLS-DA class prediction for HS-GC-IMS. Classes are from top to bottom respectively class1, class2 and class3. In each figure on the left of the vertical line the calibration set samples, on the right the prediction set samples.

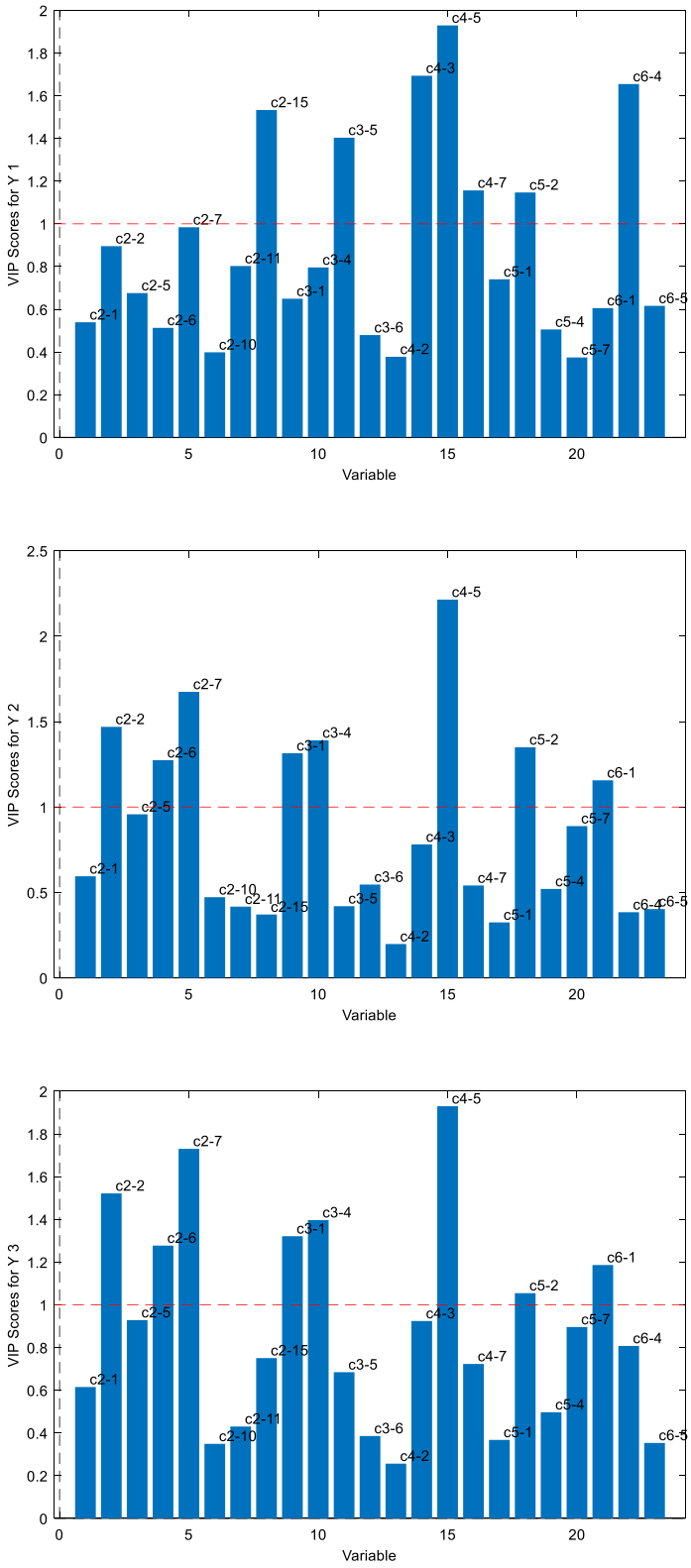


Figure 5-7. PLS-DA model based on HS-GC-IMS data. VIP scores of the variables for each class. Labels indicates the interval and the relative component number.

5.2.2 GC-FID e-nose data

The whole chromatogram obtained with the MXT5 column was considered. The chromatograms were normalized for the respective internal standard, then aligned on retention time using icoshift algorithm (see chapter 3). The resulting chromatograms were used in an explorative PCA (data not reported for sake of brevity) to have an overview on the distribution of samples.

Again PLS-DA, after calibration and validation samples splitting, was used to inspect pesto distinction by classes. To estimate the correct number of Latent Variables of PLS-DA, cross-validation was performed with a venetian blind scheme using 10 splits. Four Latent Variables were selected.

In Figure 5-8 is reported the scores plot for the first two LVs. It is possible to observe a good separation of the samples, especially for class 2, but also for classes 1 and 3. Moreover, it is possible to observe how the test samples (indicated by a T) are close to the respective classes.

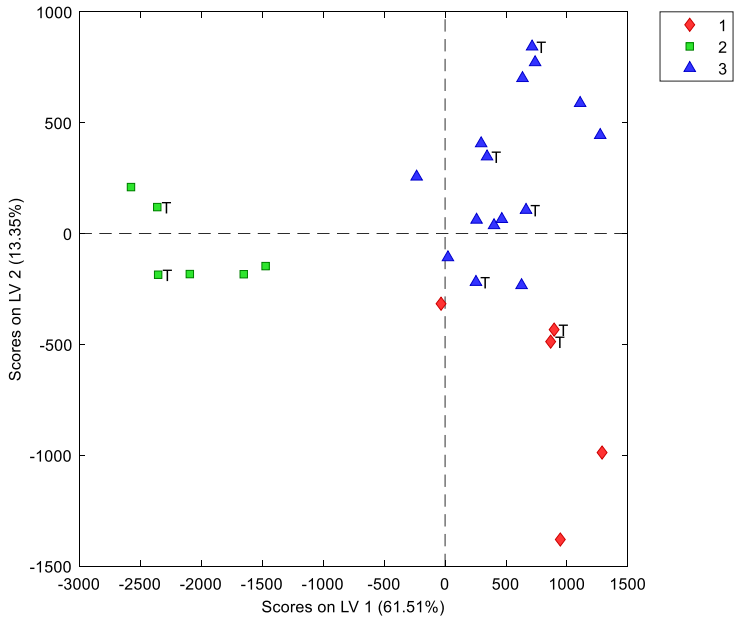


Figure 5-8. PLS-DA for GC-FID e-nose data. Samples are coloured by classes, while the test set samples are indicated with the letter T.

The confusion matrix in cross validation and in prediction, reported in Table 5-4 and Table 5-5 respectively indicates that all classes are correctly predicted. The same could be observed in Figure 5-9, where are reported the values of the predicted Y-values vs. N° of samples with the respective thresholds.

Table 5-4. Confusion matrix reporting the number of samples recognized in Cross Validation for each class (PLS-DA model obtained by GC-FID e-nose data)

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	2	0	0
Predicted as 2	0	4	0
Predicted as 3	1	0	11
unassigned	0	0	0

Table 5-5. Confusion matrix reporting the number of samples recognized in Prediction (test set) for each class (PLS-DA model obtained by GC-FID e-nose data)

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	2	0	0
Predicted as 2	0	2	0
Predicted as 3	0	0	4
unassigned	0	0	0

The VIP scores for each class, reported in Figure 5-10, give information on which part of the chromatogram is relevant in separating the pesto classes.

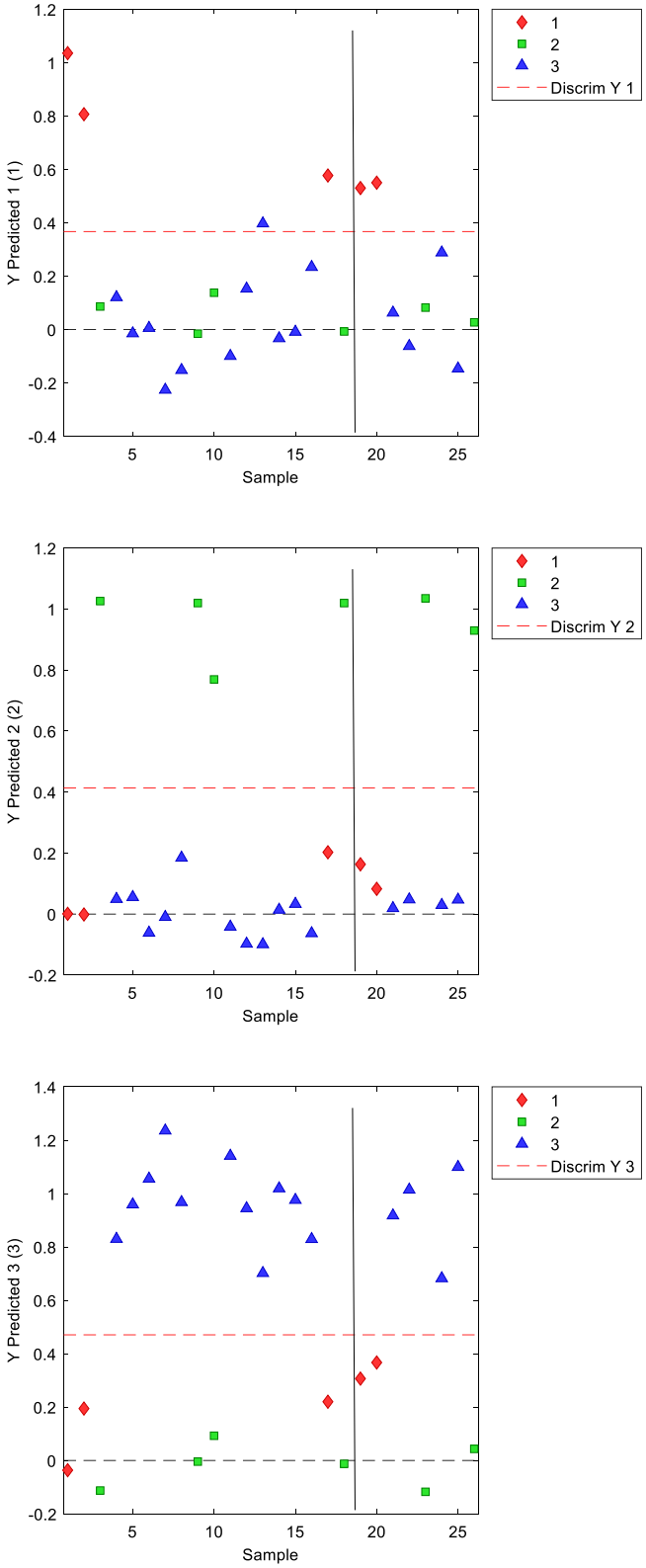


Figure 5-9. PLS-DA based on GC-FID e-nose data. Plot of predicted Ys for each class. Samples are coloured per classes, while test set samples are on the right of the vertical line in each figure. The red lines represent the class membership threshold.

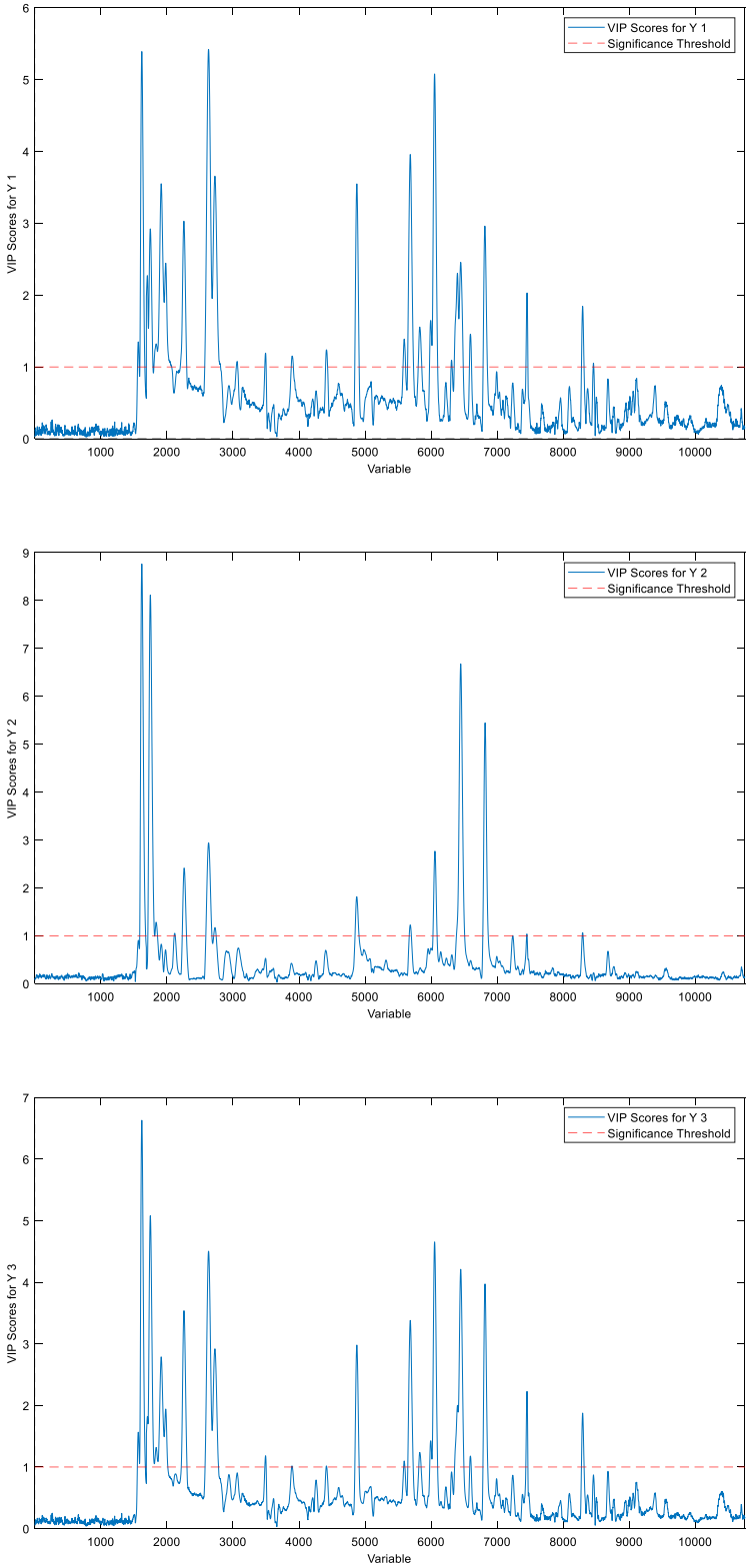


Figure 5-10. PLS_DA for GC-FID e-nose. VIP scores for the three predicted classes.

By looking at the VIP scores for each class, is not so easy to depict which peaks are responsible for the classes' differences, so to see which are the most discriminant feature, and to highlight which can be worth of quantification through analysis with standards. Thus, a parsimonious variable selection routine, i.e. CovSel, was applied to recover the most discriminant peaks.

The maximum number of variables to select was set to 20 and after inspecting the plot of explained Y-variance vs. number of selected variables, 8 variables were retained, that correspond to the peaks reported in Figure 5-11. It is interesting to observe that these variables, selected by CovSel correspond to peaks which were putatively identified (name reported in the figure) as molecules that for the majority were already known as important for the pesto aroma. The other peaks indicated by CovSel will be further investigated and could be related with other ingredients present in pesto.

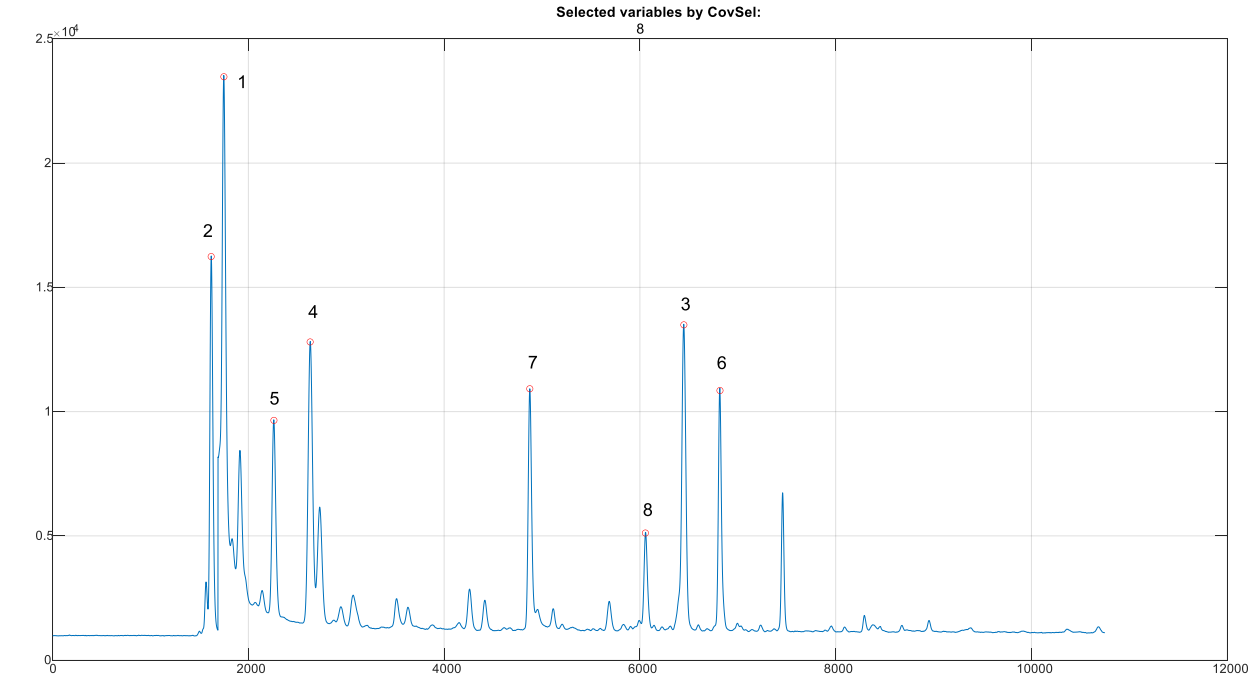


Figure 5-11. Heracles II (HS-GC-FID). Chromatogram of MXT5 column. In red the 8 variables selected by CovSel. Some of them correspond to peaks of already identified as pesto aroma molecules: hexanal (5), 2-butenal (4), 2-hexenal (7), myrcene (8), eucalyptol (3), linalool (6). The other molecules will be identified.

Using just the 8 variables selected with CovSel a new PLS-DA model was recalculated (Figure 5-12). A three LVs model in this case, was estimated according to cross-validation (venetian blind, 10 splits).

As observed the class separation is equivalent to what obtained using the whole chromatogram. This indicates also that in future investigation just the selected peaks can be used, without loss of information.

The confusion matrix confirms that the model with just 8 selected variables give the same performance in prediction (see Table 5-6 and Table 5-7).

In this case studying the VIP scores, combined with the information of the variable on the chromatogram, is possible to understand that: variable 6 and 4 (linalool and 2-butenal) are more relevant for identify class 1, variable 1, 2 (unidentified) and 7 (2-hexenal) are relevant for class 2 and variable 6 (linalool) is relevant for class 3.

This is a good example on how the chemometrics approach could give important information on the real system under study.

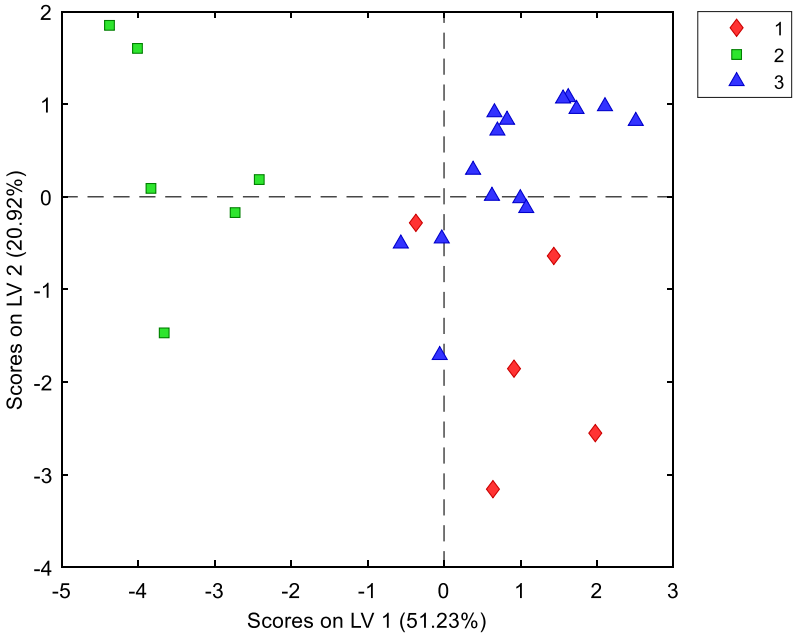


Figure 5-12. HS-GC-FID Heracles II – PLS-DA score plots with the 8 variables selected by CovSel. Different color represents the three classes.

Table 5-6. Confusion matrix reporting the number of samples recognized in Cross Validation for each class (PLS-DA model obtained by GC-FID e-nose with selected variables

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	1	0	1
Predicted as 2	1	4	0
Predicted as 3	1	0	10
unassigned	0	0	0

Table 5-7. Confusion matrix reporting the number of samples recognized in Prediction (test set) for each class (PLS-DA model obtained by GC-FID e-nose with selected variables

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	1	0	0
Predicted as 2	0	2	0
Predicted as 3	1	0	4
unassigned	0	0	0

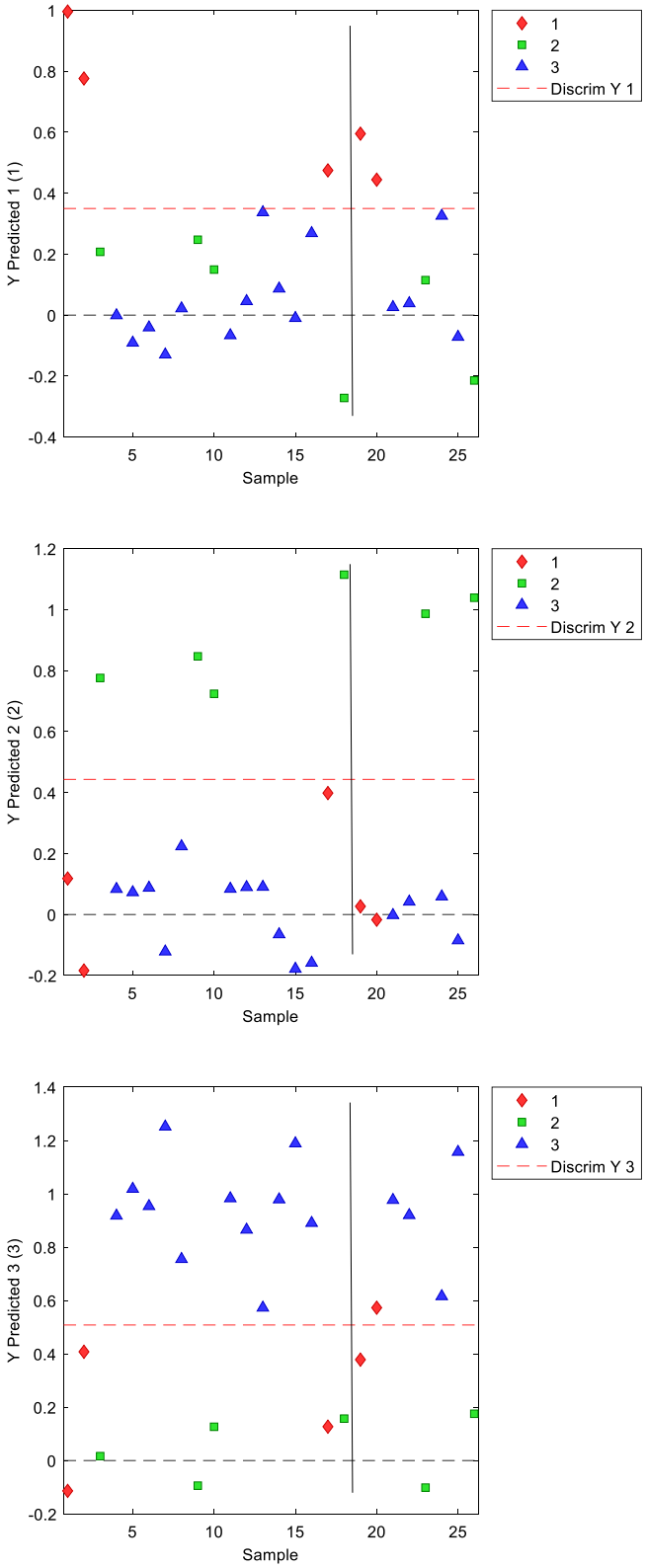


Figure 5-13. PLS-DA on GC-FID e-nose with selected variables. The colours represent the pesti classes, while test set samples are on the right of the vertical line in each figure.

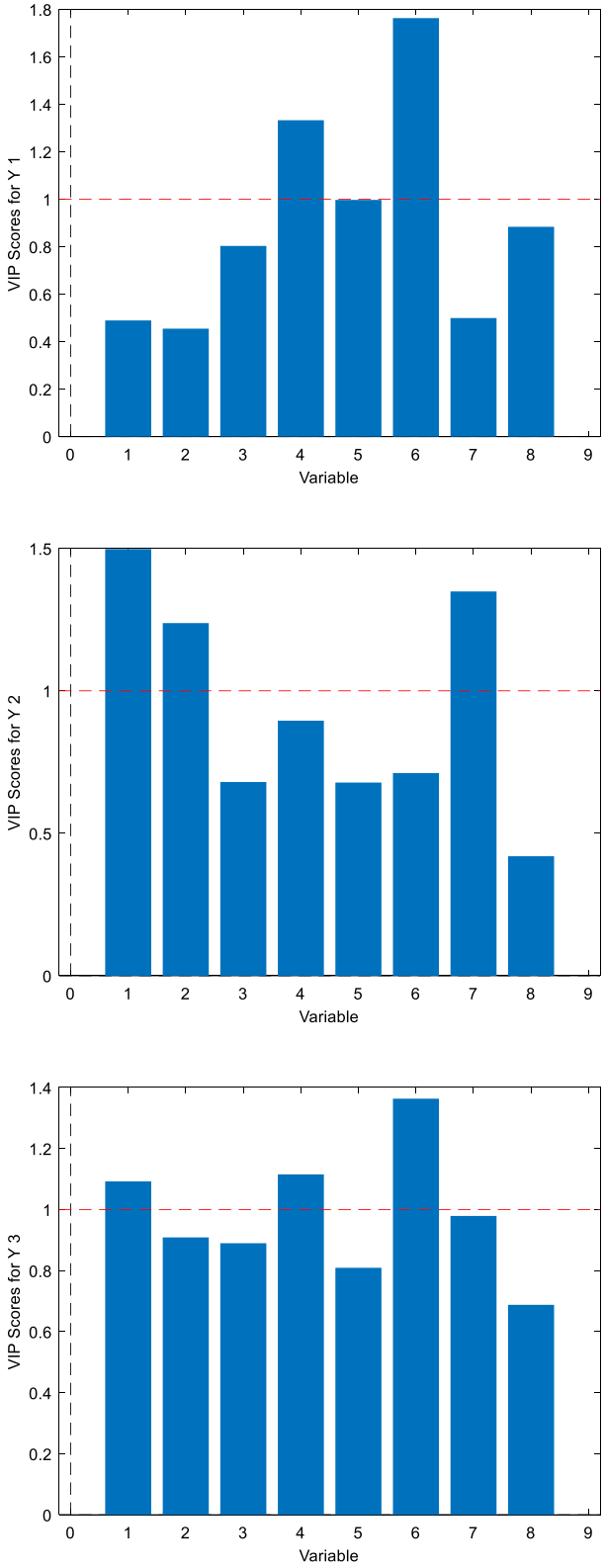


Figure 5-14. PLS_DA for GC-FID e-nose with selected variables. VIP scores for the three predicted classes.

5.2.3 NIRS Data

NIRS data (in the range 400-2500 nm) were pre-treated by Savitzky-Golay smoothing, second derivative and mean centering.

For each sample eight spectra were collected automatically by the instrument rotating the cup containing the sample, to acquire the replicates in different portion of the sample. In the preparation of the calibration and validation sets all replicates of the same sample were maintained in the same set.

The whole pre-treated spectra were used for an explorative PCA (data not reported) and then a PLS-DA was calculated. Four Latent Variables were selected according to cross validation (venetian blind, 10 splits). As in the other cases, samples were split into calibration set and validation set (144 spectra corresponding to 18 samples and 64 spectra corresponding to 8 samples respectively). As observed in Figure 5-15 PLS-DA indicated again a separation of the class 2 from the other two, that on the other hand are quite overlapped, as already observed with the other techniques.

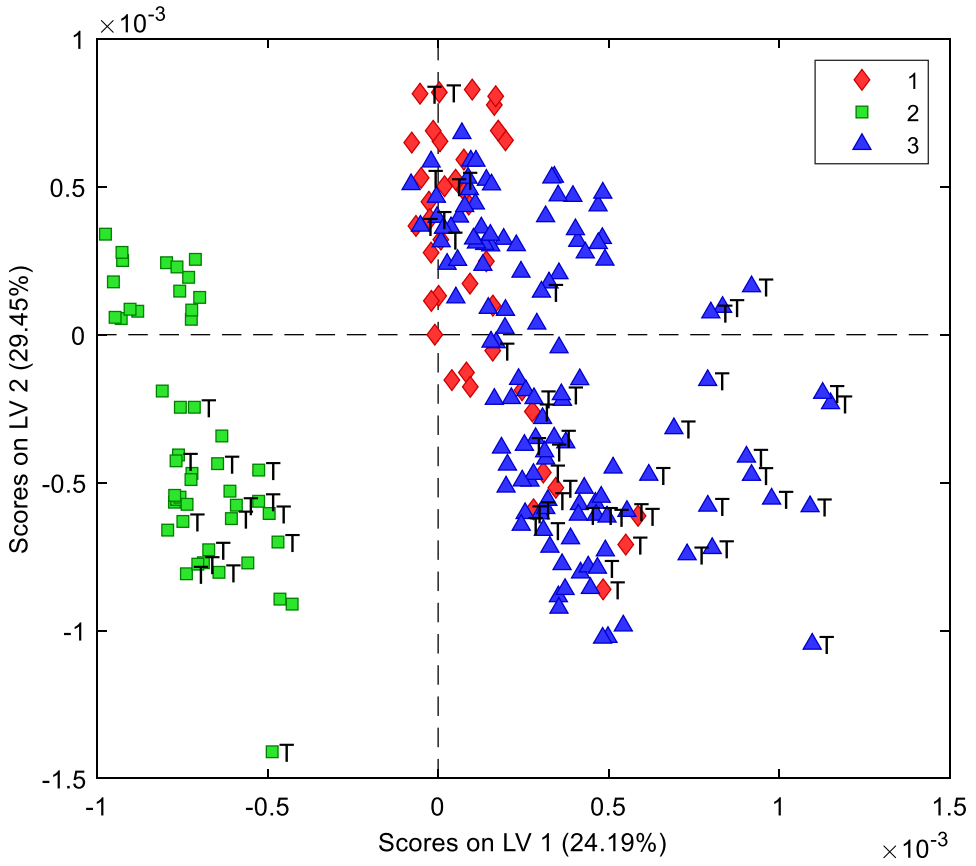


Figure 5-15. NIRS data. PLS-LDA scores plot of whole pre-treated spectra. Different color represents the three classes: Letter T indicates test set samples.

Here, the overlap between class 1 and 3 is more pronounced, as to be expected being the difference among pesto types mainly due to the aroma which can be more difficult to catch by NIRS. Anyhow some other characteristics of pesto composition may vary among classes, that could be correlated to what the other technique observes in terms of pesto aroma.

In tables Table 5-8 and Table 5-9 the confusion matrices are reported, confirming a predictive capability quite similar to the chromatographic techniques. In fact in terms of samples only one is misclassified (belonging to class 1 but predicted as class 3) and a single replicate of class 3 predicted as 1. Figure 5-16 reports the predicted Y-values for each class.

Table 5-8. Confusion matrix reporting the number of samples recognized in Cross Validation for each class (PLS-DA model obtained by NIR

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	18	0	4
Predicted as 2	0	32	0
Predicted as 3	6	0	84
unassigned	0	0	0

Table 5-9. Confusion matrix reporting the number of samples recognized in Prediction (test set) for each class (PLS-DA model obtained by NIR

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	10	0	1
Predicted as 2	0	16	0
Predicted as 3	6	0	31
unassigned	0	0	0

In this case the study of the VIPs indicates that the class 1 is more different in the visible range from 400 to 800 nm and in the NIR range from 1800 to 2000 nm, a zone that could be related to the water signals.

Also in this case, like for the GC-FID e-nose, the complexity of the original spectrum does not allow a clear interpretation.

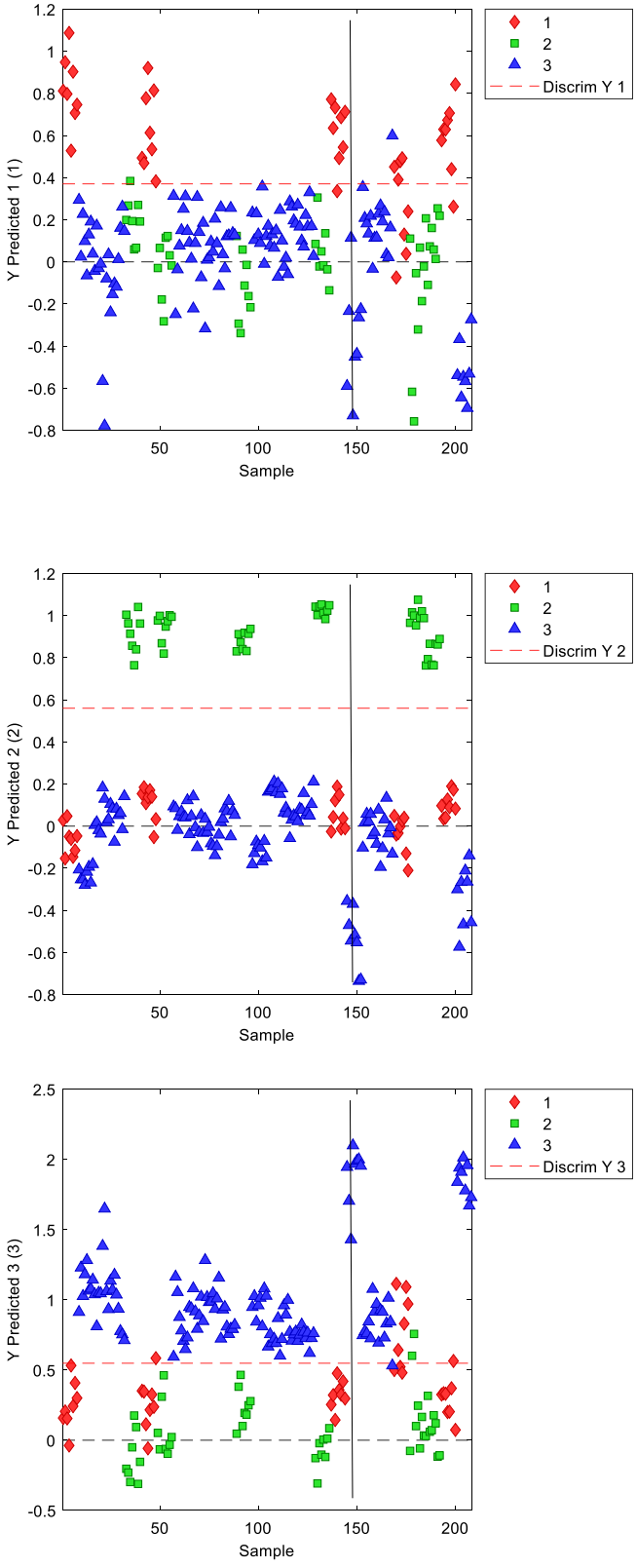


Figure 5-16. PLS-DA on NIR data. Samples predicted for the three classes from top to bottom respectively. Samples on the right of the vertical line are test set samples.

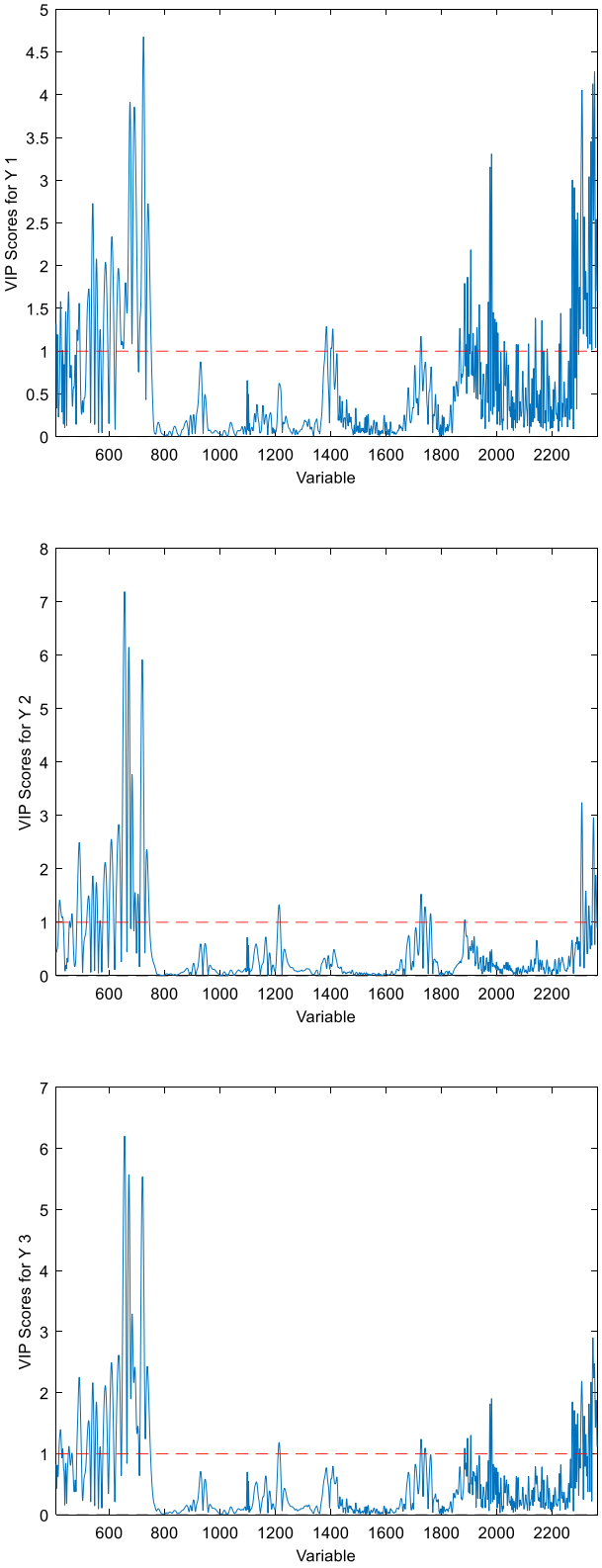


Figure 5-17. PLS-DA of NIR data. VIP scores for the three classes respectively.

Thus, CovSel was applied to highlight the most relevant spectral features. The results indicate that the discriminant spectral regions are related to colour and chlorophylls in the visible part of the spectrum, and water content and lipids in the last part of the spectrum, in NIR region (Figure 5-18).

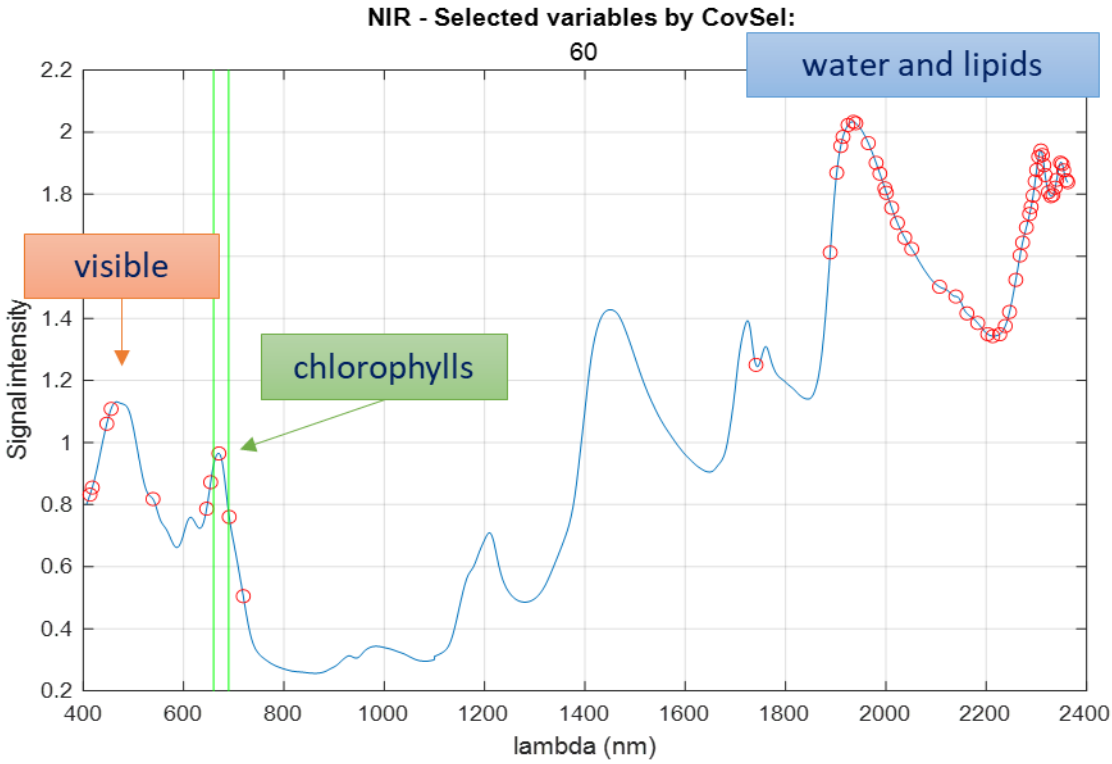


Figure 5-18. NIRS pesto spectrum. Red circles are the 60 variables indicates by CovSel to better separate the three pesti classes.

Again, a PLS-DA (5 LVs according to venetian blind CV, 10 splits) was calculated with just the variables selected by CovSel.

The results are slightly worse than the one obtained by the whole spectrum model, however coherent with them.

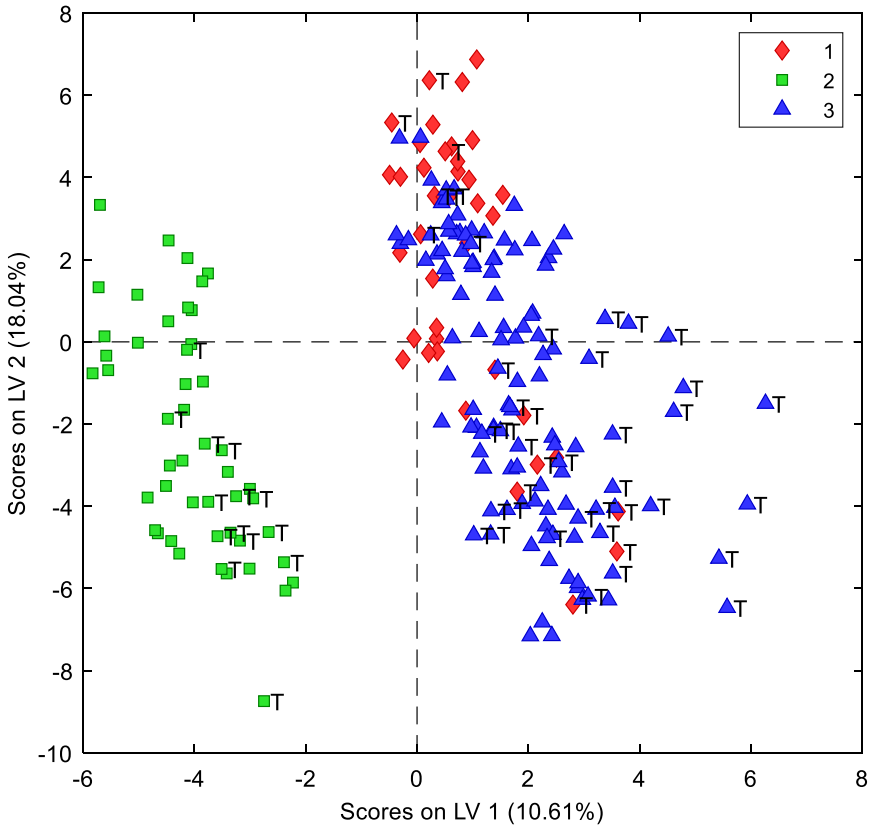


Figure 5-19. PLS-DA on NIR with selected 60 variables. Latent Variable 1 and 2 are reported. Samples with le letter T are the test set.

Observing the confusion matrixes is possible to observe that also in this case the classes membership prediction is still quite good, as could be already observed in Figure 5-20 that reports the predicted samples for each class.

The observation of the VIP scores for each class suggest that the main differences are in the visible part of the NIR spectrum (the colour of the pesto), but also in the zone of the water and lipids.

Table 5-10. Confusion matrix reporting the number of samples recognized in Cross Validation for each class (PLS-DA model obtained by NIR with selected variables

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	21	0	6
Predicted as 2	0	32	0
Predicted as 3	3	0	82
unassigned	0	0	0

Table 5-11. Confusion matrix reporting the number of samples recognized in Prediction (test set) for each class (PLS-DA model obtained by NIR

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	7	0	0
Predicted as 2	0	16	0
Predicted as 3	9	0	32
unassigned	0	0	0

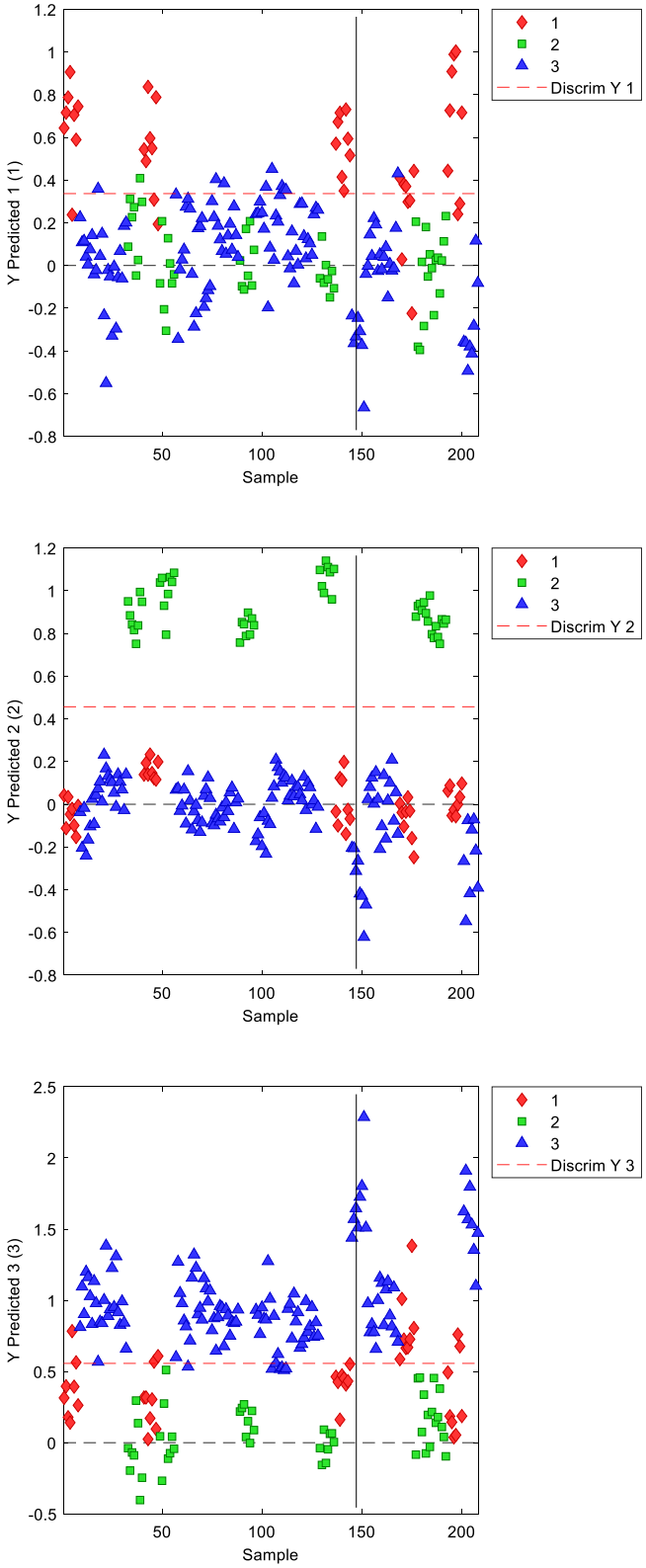


Figure 5-20. PLS-DA for NIR with 60 selected variables with CovSel. Predicted samples for each class are reported.

5.2.4 Data fusion

To combine all the three datasets a multiblock approach was considered. The NIRS block holds the whole spectra (averaged by replicates), pre-processed as described in section 3.2.2. Analogously the GC-FID e-nose data blocks hold the whole chromatograms pre-processed as described in section 3.2.1.

The GC-IMS data block was assembled by considering the peaks areas of the 31 MCR components.

Prior to multiblock data analysis samples were split into 18 calibration samples and 8 validation samples, to gather the model performance in prediction. Then block scaling and mean centring was applied to have fair contribution from each block when applying multiblock PLS-DA.

Six Latent Variables were selected (according to CV, venetian blind, 6 splits).

As observed in Figure 5-21 the class 2 continue to be separated properly form the other two. Respect to the single elaborations for each technique in this case also classes 1 and 3 seems to be less overlapped.

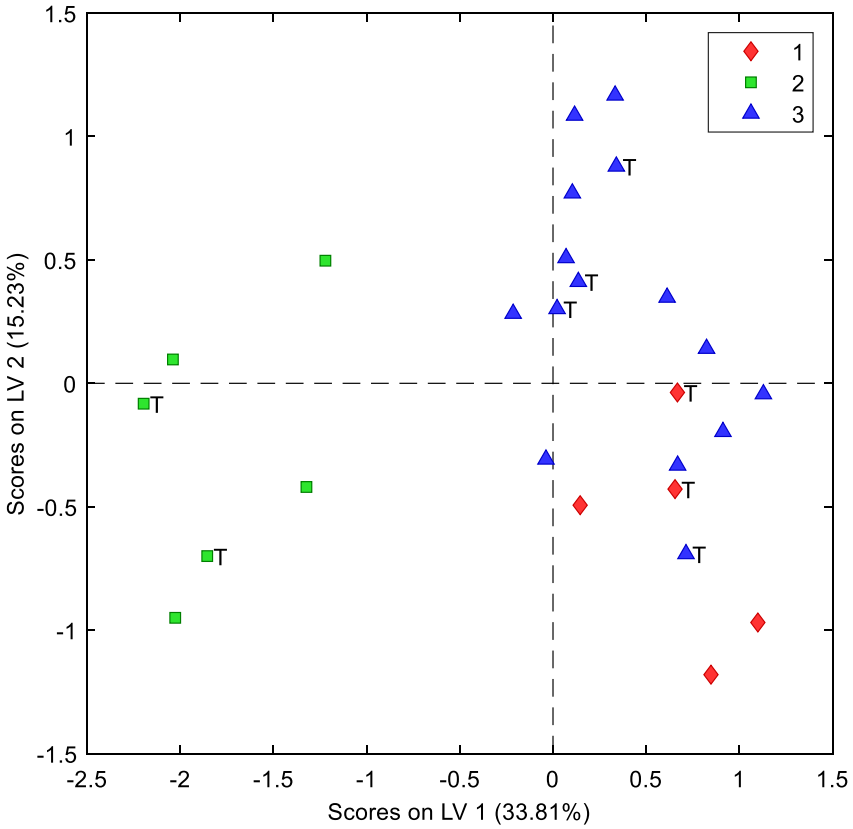


Figure 5-21. PLS-DA score plots on Low Level Data Fusion dataset without variable selection. Different colors represent the three classes, while samples with letter T represents the test set samples.

The confusion matrix reported in Table 5-12 and Table 5-13 confirm the good prediction of the three classes.

Same consideration cold be done observing the Figure 5-22 where are reported the predicted Y-values for each class.

Table 5-12. Confusion matrix reporting the number of samples recognized in Cross Validation for each class (PLS-DA model obtained by low level data fusion

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	3	0	2
Predicted as 2	0	4	0
Predicted as 3	0	0	9
unassigned	0	0	0

Table 5-13. Confusion matrix reporting the number of samples recognized in Prediction (test set) for each class (PLS-DA model obtained by low level data fusion

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	2	0	1
Predicted as 2	0	2	0
Predicted as 3	0	0	3
unassigned	0	0	0

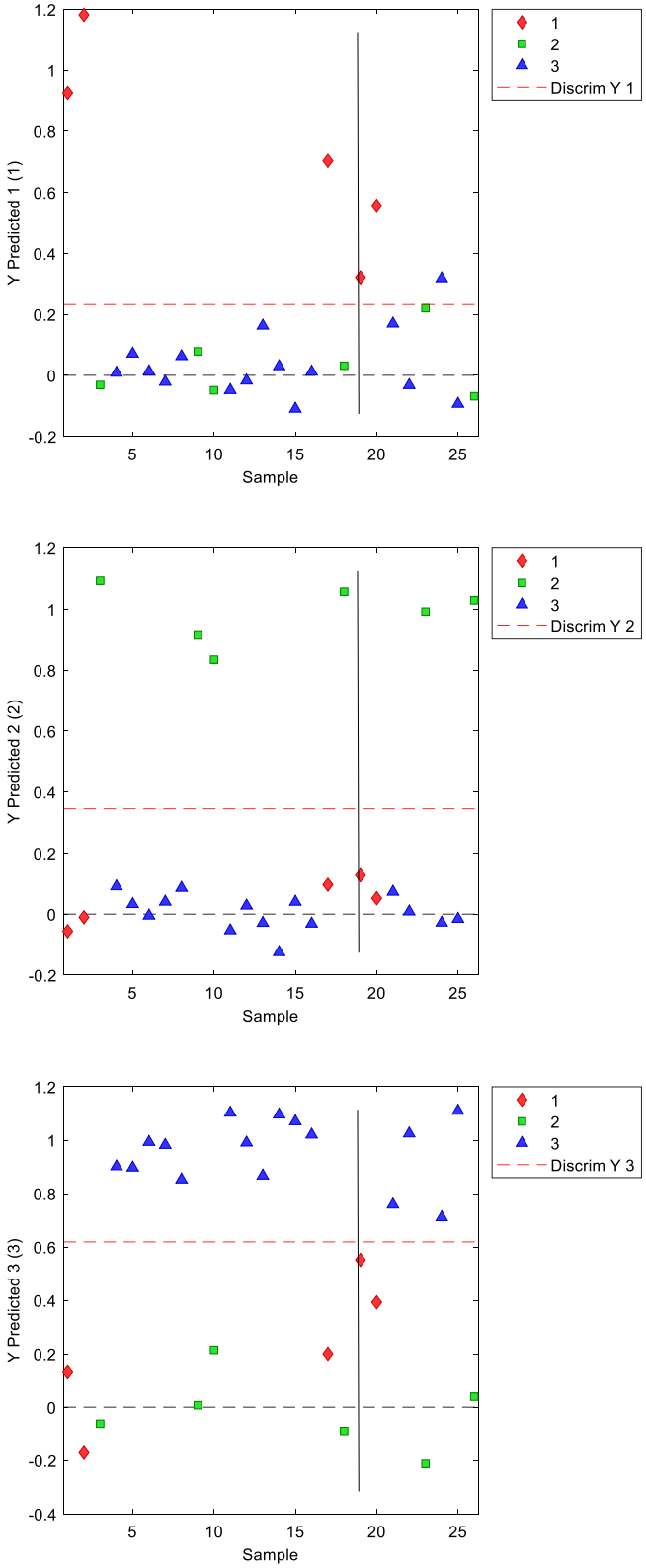


Figure 5-22. PLS-DA on low level data fusion. Samples predicted for the three classes from top to bottom respectively. Samples on the left of the vertical line are calibration set samples, while samples on the right of the vertical line are test set samples.

The VIP scores reported in Figure 5-23, coloured by block shows that the GC-IMS seems to be more relevant in order to separate the pesto classes.

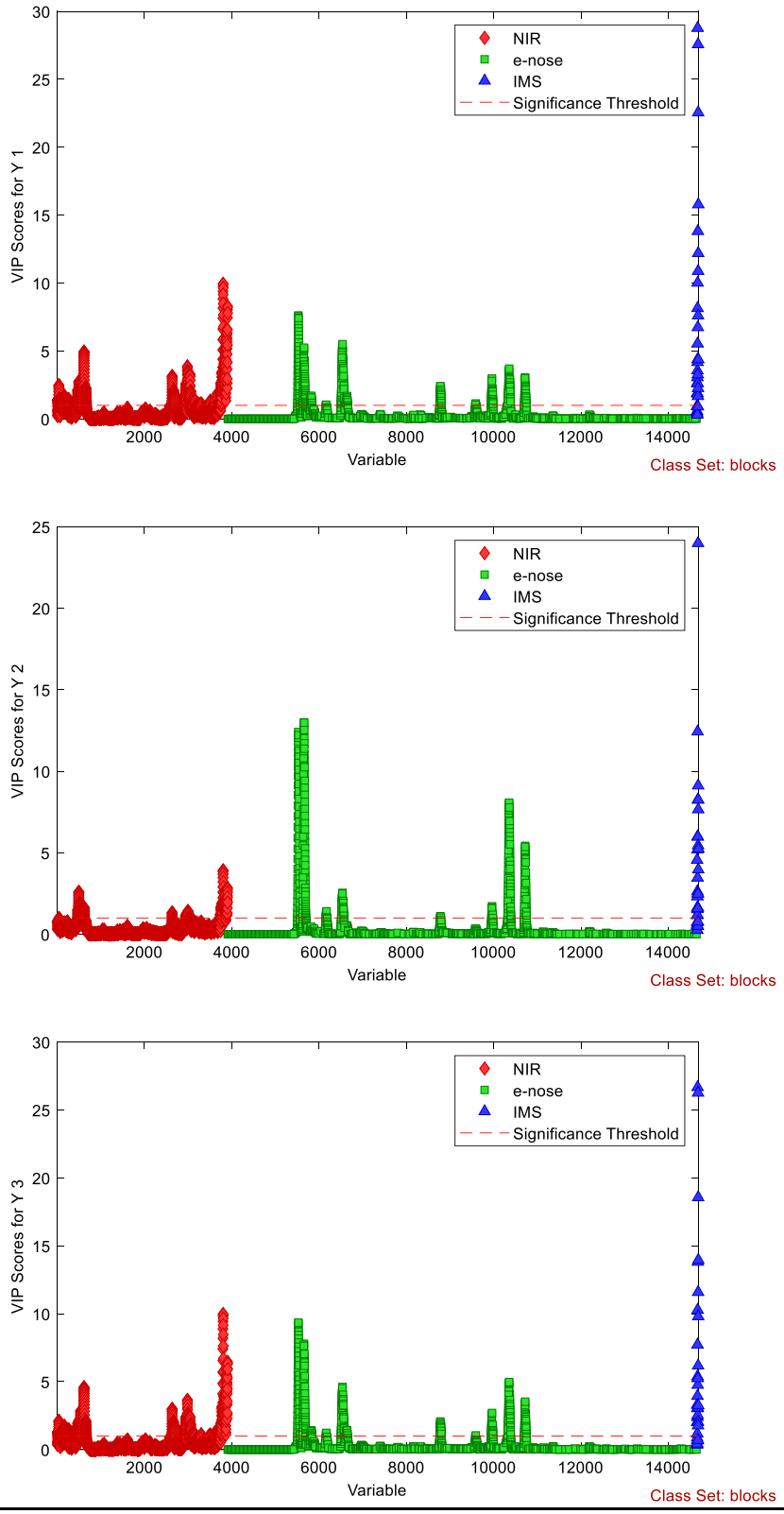


Figure 5-23. VIP scores of PLS-DA for low level data fusion. Different colours indicated the three blocks of data.

Also in this case to further interpret the role of the different blocks/variables the CovSel algorithm has been applied on the low-level fused dataset. The initial number of selected variables was set to 20, then, observing the cumulative variance plot (Figure 5-24) the variable number chosen was 10.

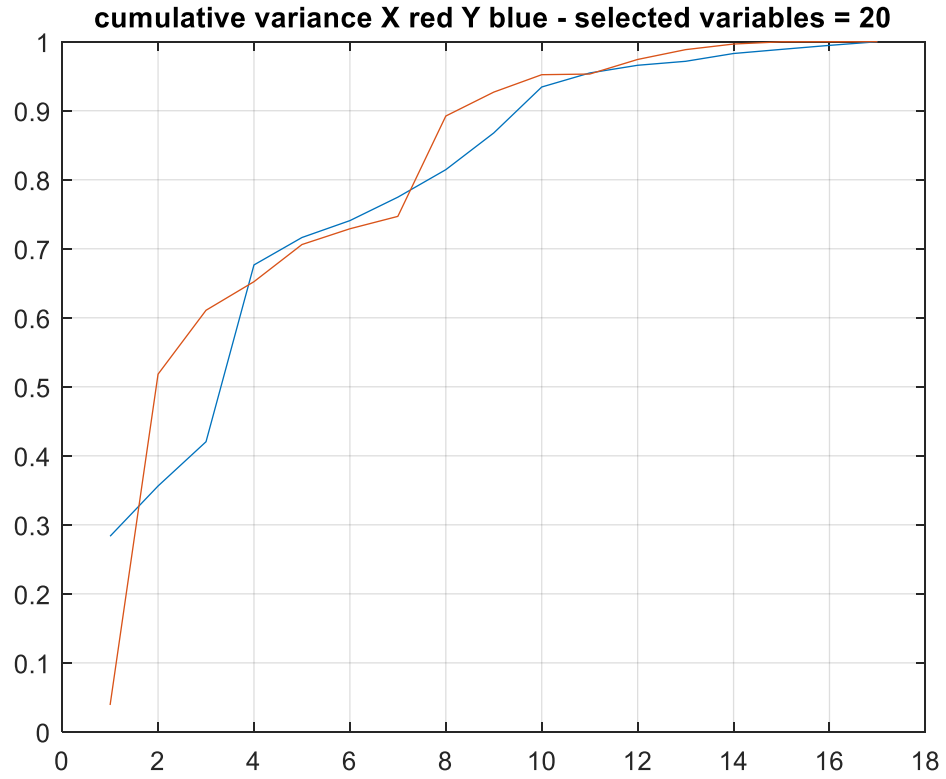


Figure 5-24. Cumulative variance for X and Y for CovSel applied on low-level data fusion.

The selected variable, belong: three to the NIR block, one to the GC-FID e-nose block and six to the GC-IMS block. With these ten variables a new PLS-DA model was built with four latent variables (selected by cross validation with a venetian blind scheme with 10 splits). However, this reduced model performs poorly, maintaining some predictive capability just for the class 2 (see Figure 5-25 and confusion matrixes in Table 5-14 and Table 5-15).

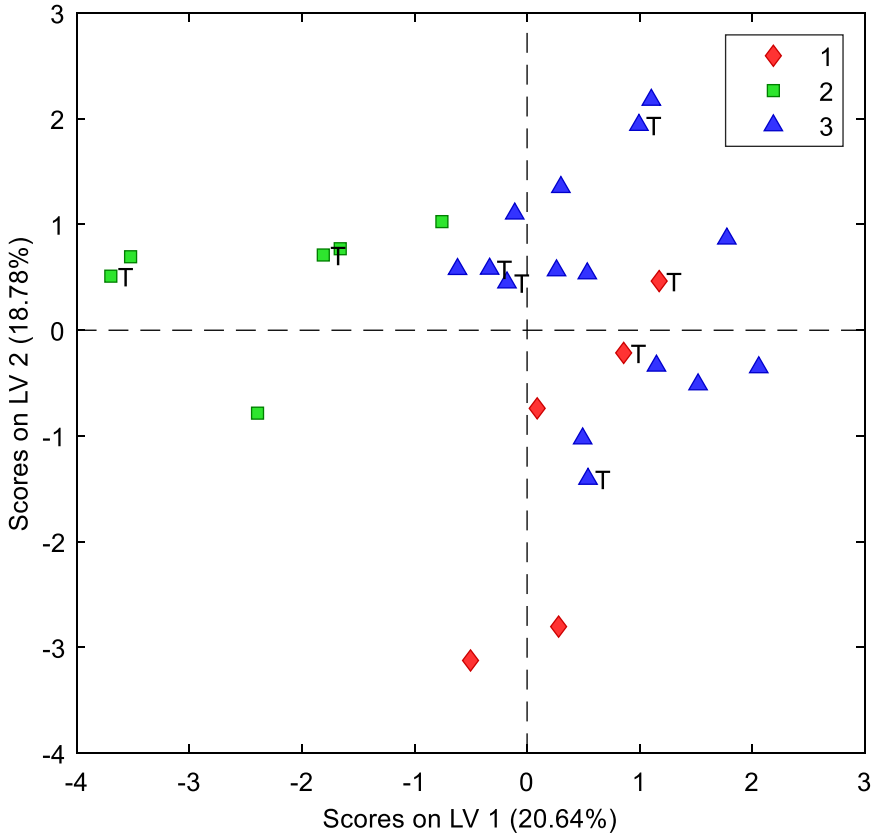


Figure 5-25. PLS-DA with variables selected by CovSel on the low-level fused dataset. Letter T indicates test set samples.

Table 5-14. Confusion matrix reporting the number of samples recognized in Cross Validation for each class (PLS-DA model obtained by low level data fusion, after CovSel with 10 variables.

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	2	1	0
Predicted as 2	1	3	1
Predicted as 3	0	0	10
unassigned	0	0	0

Table 5-15. Confusion matrix reporting the number of samples recognized in Prediction (test set) for each class (PLS-DA model obtained by low level data fusion after CovSel with 10 variables.

	Actual class 1	**Actual class 2**	**Actual class 3**
Predicted as 1	0	0	0
Predicted as 2	0	2	0
Predicted as 3	2	0	4
unassigned	0	0	0

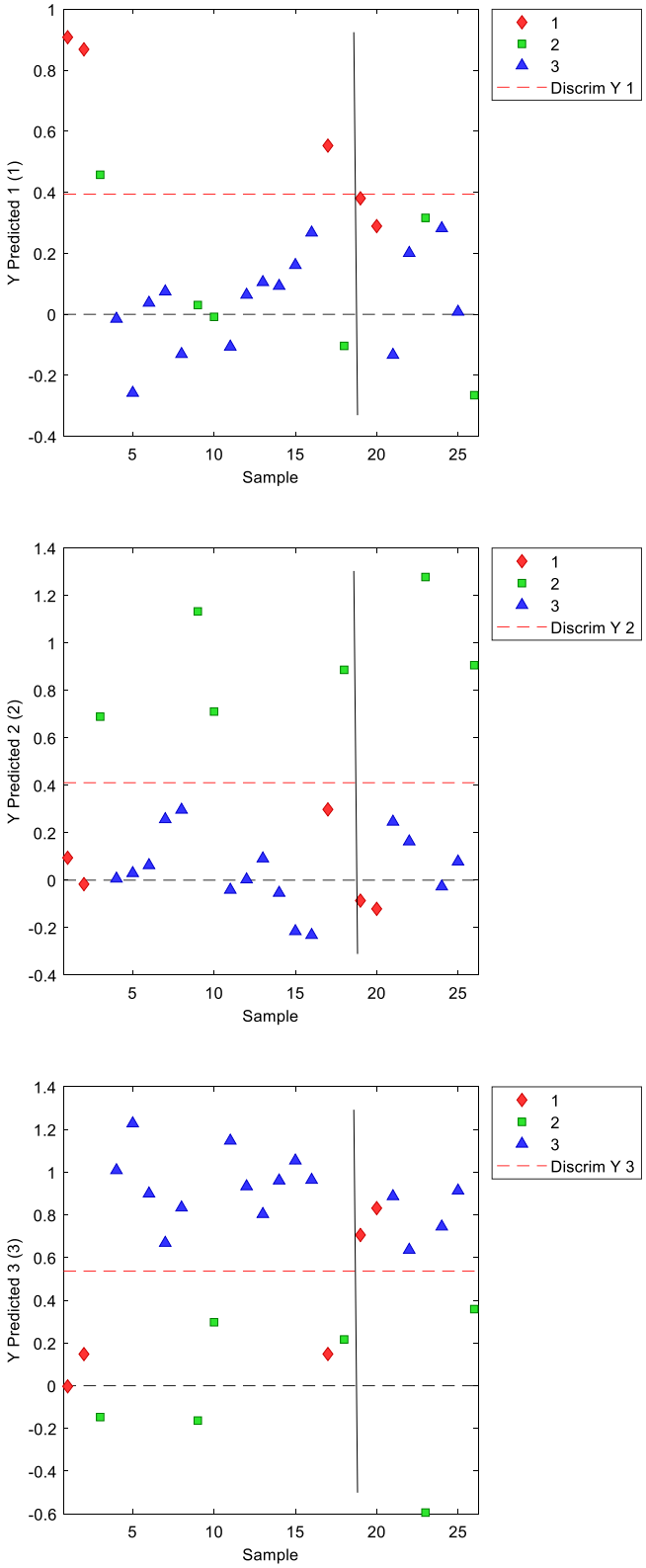


Figure 5-26. Sampe prediction done with model on CovSel selected variables on low-level fused dataset.

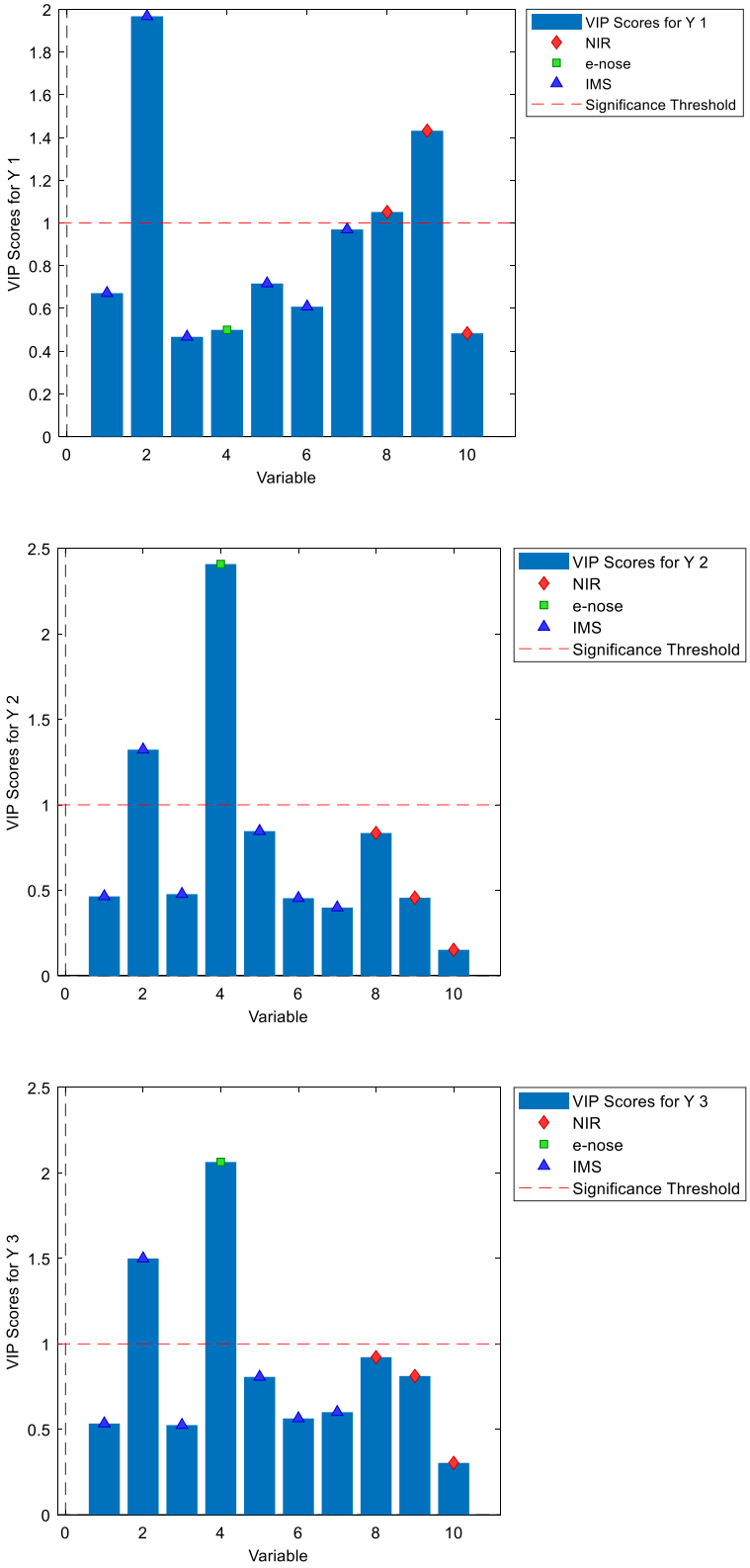


Figure 5-27. VIP scores of model with variable selected by CovSel on low-level-fused dataset.

5.3 Conclusions

All the three techniques, HS-GC-IMS, GC-FID e-nose and NIRS, singularly evaluated were able to perfectly discriminate class 2 from the others two, and satisfactorily class 1 and 3.

In the first two cases the separation could be attributed to the aromatic pattern, while in the case of NIRS, differences were due to pesto colour, chlorophyll, water, and lipids contents.

The GC-FID e-nose shows the better performances in separating the three pesto classes. Class 2 is very well separated, but also the other two classes are quite distinguished.

The use of the whole chromatogram by the GC-FID e-nose is appealing since it catches all the information of volatile molecules, while being rapid and easy to operate. The same good performance remains after the variable selection done with CovSel, with just a moderate loss.

HS-GC-IMS, that similarly to GC-FID e-nose works on volatile molecules, separates in a proper way the class two but less well the other two. In this case the more complex routine to extract information from the 3D chromatogram could have influenced the final performances. For this technique, due to the low number of variables obtained from the MCR-ALS variable selection has not been tested.

The NIR, also gave a satisfactory performance in this case not only the volatile molecule profile, but also colour, chlorophylls, water, and lipids content contribute to classes separation.

Thus, the NIRS technique could be also used to characterise the pesto classes, instead of more complex and time-consuming techniques.

After variable selection the performance decreased, while remaining acceptable, suggesting that the main differences lie in the chlorophyll, water and lipids content. This could be an indication for eventually adopting this simpler and cheaper analytical technique for fast screening.

The combination of the information of the three techniques did not give in this case a significant increase in discrimination performance, nonetheless, providing reduced classes overlap, contrary to single technique models, here when using only the Covsel selected features the predictive performance decreased.

However, the data fusion approach gives the possibility to better understand which analytical technique is more useful for the class characterization of pesto samples. In this specific case the relevant information is that NIRS, more easy, flexible, and exportable technique, can successfully characterize pesto, with a potential application in an industrial environment.

---

1 Jurado-Campos N, Martín-Gómez A, Saavedra D, Arce L., Usage considerations for headspace-gas chromatography-ion mobility spectrometry as a suitable technique for qualitative analysis in a routine lab, Journal of Chromatography A, 2021;1640: 461937 https://doi.org/10.1016/j.chroma.2021.461937.

2 Wiśniewska P, Śliwińska M, Namieśnik J, Wardencki W, Dymerski T. The Verification of the Usefulness of Electronic Nose Based on Ultra-Fast Gas Chromatography and Four Different Chemometric Methods for Rapid Analysis of Spirit Beverages. J Anal Methods Chem. 2016; 2016: 8763436. doi: 10.1155/2016/8763436.

3 Pellicer A, del Carmen Bravo M. Near-infrared spectroscopy: A methodology-focused review. Seminars in Fetal and Neonatal Medicine. 2011; 16(1): 42-49. https://doi.org/10.1016/j.siny.2010.05.003

4 Mäkinen MA, Anttalainen OA, Sillanpää MET. Ion Mobility Spectrometry and Its Applications in Detection of Chemical Warfare Agents Analytical Chemistry. 2010; 82(23): 9594-9600 doi: 10.1021/ac100931n

5 Dodds JN, Baker ES. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. J Am Soc Mass Spectrom. 2019 30(11): 2185-2195. doi: 10.1007/s13361-019-02288-2.

6 Eilers PHC, Boelens HFM. Baseline correction with asymmetric least squares smoothing. Leiden University Medical Centre Report. 2005; 1(1): 5.

7 Cavanna D, Zanardi S, Dall'Asta C, Suman M. Ion mobility spectrometry coupled to gas chromatography: A rapid tool to assess eggs freshness, Food Chemistry. 2019; 271:691-696. https://doi.org/10.1016/j.foodchem.2018.07.204.

8 Wang S, Chen H, Sun BRecent progress in food flavor analysis using gas chromatography–ion mobility spectrometry (GC–IMS), Food Chemistry. 2020;315: 126158 https://doi.org/10.1016/j.foodchem.2019.126158.

9 Christmann J, Rohn S, and Weller P. Gc-ims-tools–A new Python package for chemometric analysis of GC–IMS data. Food Chemistry. 2022; 394: 133476.

10 Capitain C and Weller P. Non-targeted screening approaches for profiling of volatile organic compounds based on gas chromatography-ion mobility spectroscopy (GC-IMS) and machine learning. Molecules. 2021; 26(18): 5457.

11 Brendel R, Rohn S, Weller Pand Mannheim H. GC-IMS and machine learning for the rapid differentiation of complex isomeric mixtures of erpenes. Lebensmittelchemie. 2023; 77: S2-008.

12 Szymańska E, Davies AN, and Buydens LM. Chemometrics for ion mobility spectrometry data: recent advances and future prospects. Analyst. 2016; 141(20): 5689-5708.

13 Oller-Moreno S, Singla-Buxarrais G, Jiménez-Soto J M, Pardo A, Garrido-Delgado R, Arce L, et al. Sliding window multi-curve resolution: application to gas chromatography–ion mobility spectrometry. Sensors and Actuators B: Chemical. 2015;217: 13-21.

14 Parastar H. Chapter 6 - Multivariate Curve Resolution Methods for Qualitative and Quantitative Analysis in Analytical Chemistry. Data Handling in Science and Technology2015; 29: 293-345, https://doi.org/10.1016/B978-0-444-63527-3.00006-0.

15 Ortiz-Villanueva E, Benavente F, Piña B, Sanz-Nebot V, Tauler R, Jaumot J. Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data, Analytica Chimica Acta. 2017; 978: 10-23. https://doi.org/10.1016/j.aca.2017.04.049.

16 Brendel R, Schwolow S, Rohn S, Weller P. Comparison of PLSR, MCR-ALS and Kernel-PLSR for the quantification of allergenic fragrance compounds in complex cosmetic products based on nonlinear 2D GC-IMS data, Chemometrics and Intelligent Laboratory Systems. 2020; 205: 104128., https://doi.org/10.1016/j.chemolab.2020.104128.

6IMAGING APPLICATIONS FROM RGB TO HYPERSPECTRAL IMAGES

It has been already underlined how the basil characteristics may impact on the quality of the final product pesto. In particular, the basil characteristics, in terms of flavour, colour, fibre and water content, have a heavy impact on the finished product quality. Moreover, different chemotype of Genovese basil and its method of cultivation could give differences in term of flavour or colour.

One of the critical characteristics of the "*Pesto alla Genovese*" is its smooth structure due to the emulsion of oil in aqueous phase. A consistent oil release could indicate a loss of stability in the emulsion structure. The differences in the basil could affect the final structure of pesto for several reasons: the colour of the basil affects the colour of the final pesto; the water content and the "fibrousness" of the basil stems affects the oil:water ratio and consequently the emulsion equilibrium [1].

For these reasons the proper characterisation of basil is a crucial step.

In order to continuously monitor the incoming basil a classical RGB vision system has been installed in the very preliminary step of the pesto production line, where the basil plants enter the process. The proper elaboration of basil images will do the differences between taking just a photo and disposing of a set of precious information. Information related to the colour and the morphology of the basil plants in terms of leaves and stems, is considered highly relevant.

Hyperspectral imaging (HSI) [2] is a powerful methodology joining the possibility of describing the morphological characteristics of the sample (i.e. the image of the sample surface) to the acquisition of detailed chemical information (i.e. captured by the spectrum taken in a given wavelengths range for each single pixel of the image). In fact, with respect to classical digital images where only three (red, green, and blue) channels are acquired (RGB images), hyperspectral imaging acquires for each pixel a whole spectrum, where visible and/or near infrared range are the most common for food applications. HSI data coupled with proper data elaboration is potentially capable to give information about the chemical components and their distribution on the imaged surface. In the case of pesto, it could be very relevant to observe the different recipe ingredients/constituents distribution and to evaluate if the differences in basil origin on it.

In this Thesis, the two typologies of images were touched either on basil or pesto for evaluating different aspects and possible employment in the quality control at the plant.

6.1 RGB Vision System for on-line Basil analysis

As pointed out in the introduction to this chapter, the quality of the basil affects the quality of the pesto sauce, so in addition to the laboratory analyses, which provide an extended characterization but on fewer samples, a vision system has been implemented, at the very first step of the production line, for in-line monitoring of the basil. The system acquires RGB images while the basil is loaded on a conveyor belt, and from them some standard features are calculated by the vision system proprietary software. These are average and standard deviation of the registered intensity in small time intervals at each colour channel, plus an overall estimation of the belt area covered by basil plants.

However, there can be further refined information to extract. One issue is due to the varying illumination at the location where images are taken, which is not automatically compensated by the vision system and would require frequent recalibration of the software parameters, the conveyor belt is also covered to very different extent during production, hence giving sometimes images with very few basil plants. In addition, it would be of interest to gather an estimation of the fibrous part amount, i.e. basil stems, distinctly from the leaves amount, as well as to estimate the defects, such as black or darker spots.

With, respect to these general aims, in this study we explored different computational approaches to calculate the ratio between the leaves and the stems of the basil plants from the RGB images.

As a first attempt, we applied methods of image segmentation, by using the Otsu method [3], and different tools, present in the image analysis toolbox in Matlab, for objects detection. However, the very different illumination prevented to obtain segmentation thresholds that could satisfactorily work for all the images. The segmentation methodology was thus, only preliminary applied to annotate the ground truth for a set of calibration images, followed by manual refinement. This calibration set was then used to build classification models at pixel levels, by developing and testing three main strategies: 1) feature enhancement by applying wavelet filters + PLS-DA; 2) calculation of textural features + PLS-DA, and 3) Deep learning, by CNN net, for pixels classification.

At present, only the results of the first approach are available, while work is still in progress concerning the other two strategies, as it will be presented in the following.

6.1.1 Sampling

A prototype RGB Vision System (Sensure, Orio Al Serio, BG, Italy) [4] was installed in the Pedrignano plant to characterize the basil plants. The acquired images are usually not stored for memory constraints, however with the aim of improving the amount and significance of the extracted information some hundreds of images were manually saved during the 2021 summer production and imported in Matalb for further image processing. These RGB images have size of 1280x1020 pixels.

An example of an acquired images is shown in Figure 6-1.

Figure 6-1. RGB basil image acquired by the Vision System at the conveyor belt.

6.1.2 Feature enhancements step by WT and PLS-DA.

In the first strategy described here, the segmentation of the three main part of the image (stems, leaves, and background) was done applying the wavelet filters, then a pixel-based classification by using PLS-DA.

The wavelet transform (WT) [5] allows capturing the different frequency contributions of a signal (1D-WT) or an image (2D-WT). In the case of images WT is a good tool, not only for denoising, but as well to analyse the texture, or in other word to recover the spatial features. In short, the WT decomposes the raw image in four sub-images called Approximation (holding smooth changes, e.g. tones) and Horizontal, Vertical and Diagonal details (holding sharp, oriented changes, e.g. stripes in specific spatial direction). In this way an RGB image is decomposed in four sub-images CA, CH, CV and CD for each spectral channel (Figure 6-2). The obtained Approximation image (CA_1) can then be decomposed in turn, increasing the decomposition level, to obtain smoother and smother version of the raw image in the Approximation image at further levels. The high frequency contributions are filtered in the details sub-images (this is also referred to as multiresolution). This decomposition process is applied distinctly to each colour channel. In order, to compensate low-level distortion we used the stationary wavelet transform (SWT) implementation [6, 7, 8].

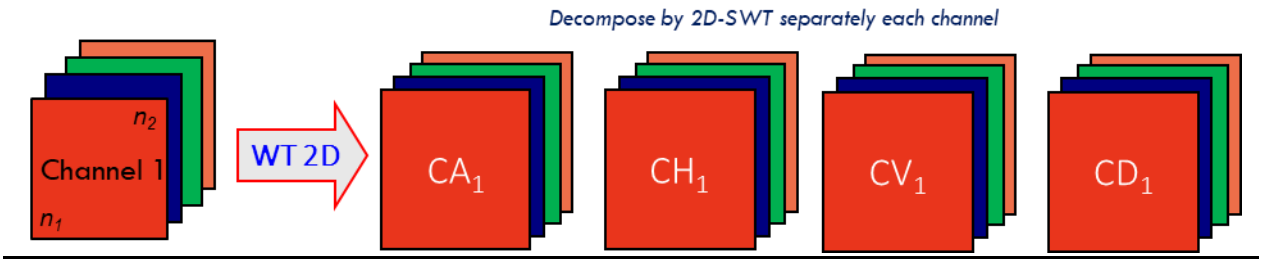


Figure 6-2. As an example, the wavelet decomposition of a four spectral channels image at the first decomposition level is shown.

From the data analysis point of view, collecting the decomposition sub-images at different levels allows setting up a multivariate data set containing enhanced information and exploiting the spatial features. For a single image, each decomposition sub-image of dimensions n1 *per* n2, is unfolded pixel-wise, obtaining a matrix of size n1*n2 rows x 4 columns. Then, the matrices corresponding to the different colour/spectral channels and decomposition levels are concatenated column-wise (Figure 6-3 left).

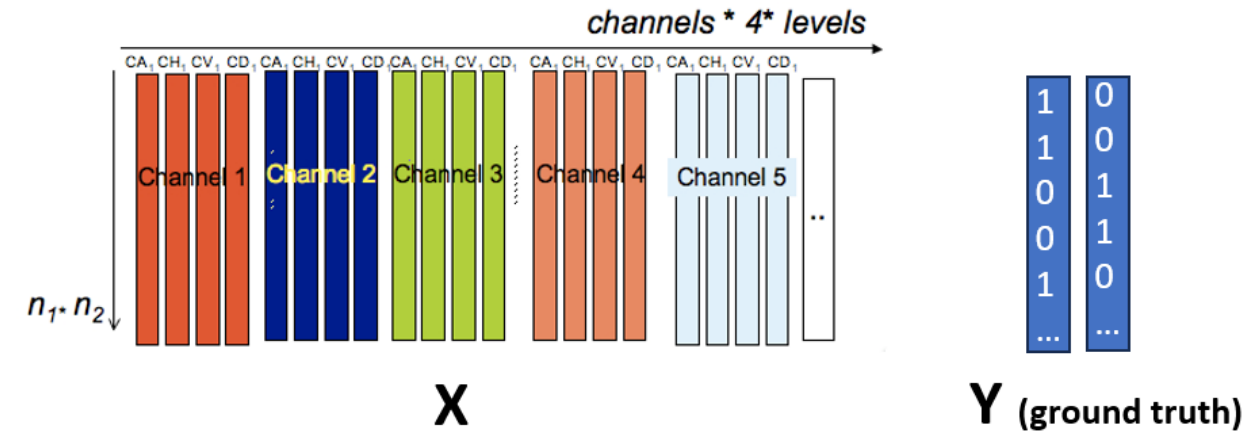


Figure 6-3. PLS-DA model 1 to predict background vs stems + leaves.

Different wavelet filters are available, for our purposes we selected the simplest, i.e. the Haar (or db1) filter and used the maximum decomposition level (namely 9) compatible with the image size.

The unfolded data set can be analysed by any multivariate method. In our case, since we wanted to achieve pixels classification PLS-DA has been applied (data was autoscaled to compensate the different scales of WT coefficients at different decomposition levels).

In particular, we used a sequential strategy: a first PLS-DA model was calculated to separate background (conveyor belt) from the rest (basil leaves + stems). Thus, the corresponding Y1 dummy matrix (Figure 6.3 Y block) was coded 1 for pixels belonging to background and 0 for the rest (basil plant, both stems and leaves); then by considering only pixels belonging to the basil plants a second PLS model is built to discriminate stems from leaves.

The calibration models have been built by using as calibration set four different images (whose ground truth has been annotated as explained before) with varying degree of conveyor belt covering and varying illumination. Because the pixels belonging to leaves are generally much more numerous than those belonging to stems, when building the second PLS-DA model the leaves class has been randomly subsampled (it is well known that any discriminant method suffers from class imbalance).

The number of PLS-DA components for both models has been selected according to minimum classification error in cross-validation (venetian blind, five splits). The classification rule adopted is to assign a pixel to the class for which the predicted Y probability is maximal.

Once the two PLS-DA models were obtained, an external set of images was predicted.

The prediction step is very fast, compatible with the on-line implementation, and is done on a new image by applying the wavelet decomposition, unfolding and concatenation to obtain the data matrix, and then applying the model 1 and model 2 in sequence:

$$[\mathbf{Y}_{background}\ \mathbf{Y}_{pred\_Stems+leaves}] = \mathbf{B}_{model1} * \mathbf{X}_{test\_all} \qquad 6\text{-}1$$

$$[\mathbf{Y}_{pred\_Stems}\ \mathbf{Y}_{pred\_Leaves}] = \mathbf{B}_{model2} * \mathbf{X}_{test\_stems+leaves} \qquad 6\text{-}2$$

Refolding the predicted class membership vector, the location of the corrected predicted pixels can be visualized (Figure 6-4).

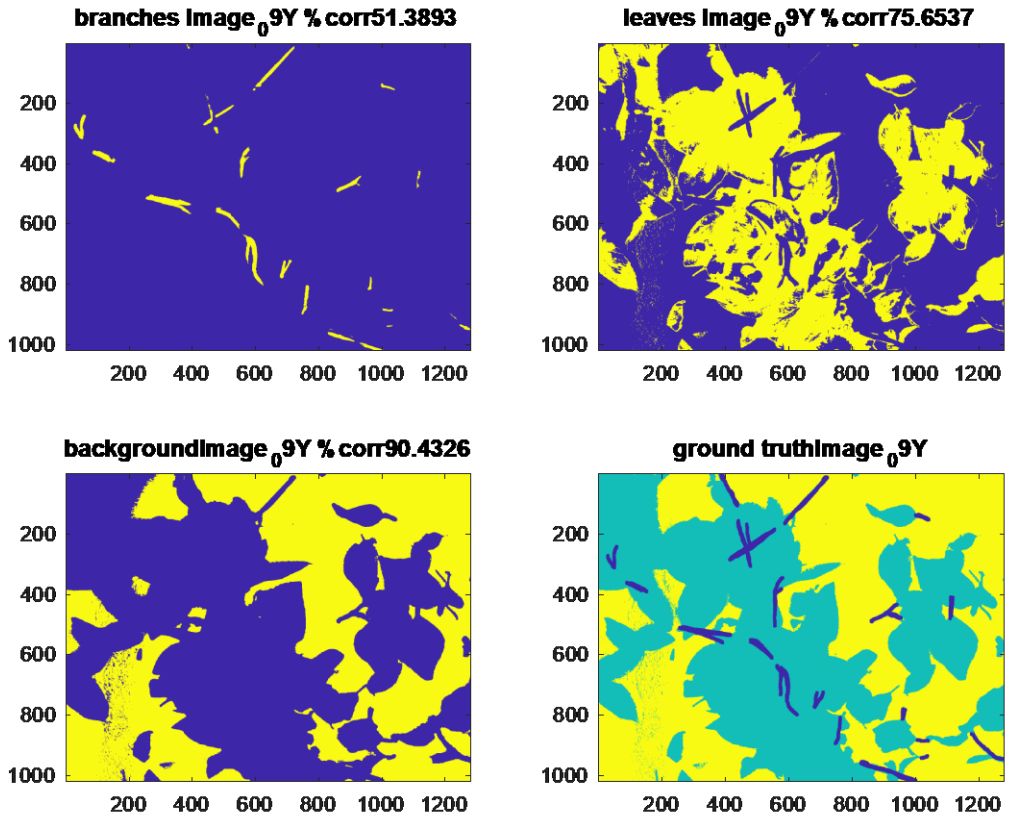


Figure 6-4. PLS-DA predictions for one of the test images, i.e. number 9. The plots show the refolded predicted class membership and the ground truth image (bottom right). Top left stems image, top right leaves, bottom left, background.

6.1.2.1 Results

In Table 6.1 are shown the results obtained for the test images, while for the calibration set considering the four images altogether the following percentages of corrected classified pixels were obtained: 69%, 86%, and 90% for stems, leaves, and background, respectively.

Table 6.1. Percent of correct pixels classification in prediction for the test set images.

Image number	Stems % correct pixels classification	Leaves % correct pixels classification	Background % correct pixels classification
9	51,4	75,7	90,4
10	55,0	66,5	94,5
11	74,5	48,3	98,7
12	54,2	83,0	32,8
13	51,2	89,1	68,1
14	48,7	88,8	25,5
15	45,5	88,5	28,1
16	53,5	81,2	62,0
42	31,8	82,7	82,9
43	51,9	75,5	92,4
44	42,4	74,5	91,1
45	45,3	72,4	96,9
46	54,2	69,6	92,4
97	55,1	74,5	91,9
98	57,8	67,9	98,2
99	72,2	63,0	95,0

As it is possible to observe, the percentage of correct prediction is good for most of the samples, with very few exceptions. The worst predictions obtained in the case of background (images number 12, 14, 15) correspond to images where the background pixels are a minority, and it must be considered that for PLS-DA model 1, since in general background was proportional to leaves, correction for imbalance was not applied. Stems in general show lower correct predictions percentage, the reason could be that they are often of the same colour of the nerves of the leaves and thus share some similarity with leaves and can be confused. Nonetheless most of them are depicted, e.g. see Figure 6-5 where predictions for image number 44 (one of the lowest correct %) are shown.

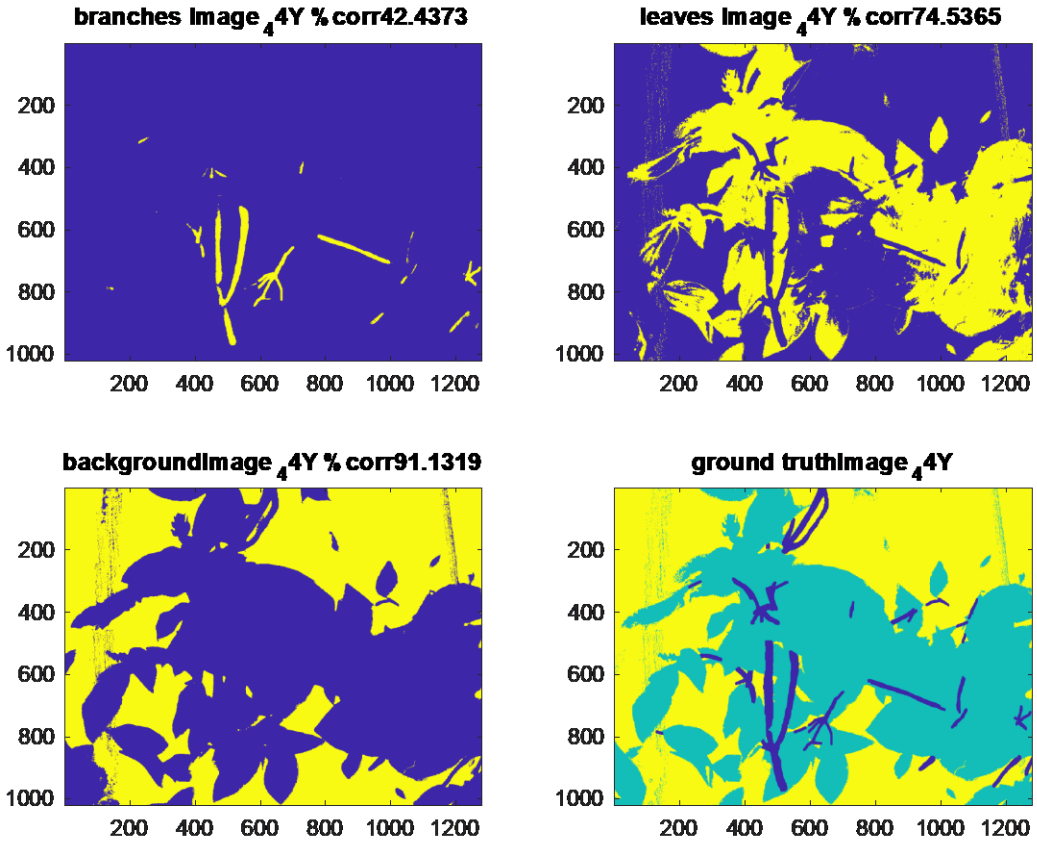


Figure 6-5. Predicted pixels memberships for test image of sample 44. Top left stems image, top right leaves, bottom left, background, and bottom right the ground truth image.

Work is in progress to refine the obtained model and to compare with the other strategies.

In Table 6.3, a preliminary classification obtained by Deep Learning architecture, described in the chapter 3 is reported. The calibration images used for the learning phase were 104 in this case.

Table 6.3. Percent of correct pixels classification in prediction for the test set images obtained by a DeepL architecture.

Image number	Stems % correct pixels classification	Leaves % correct pixels classification	Background % correct pixels classification
9	19,66	83,81	98,84
10	2,01	41,60	51,95
11	0,00	13,36	83,34
12	1,91	66,31	33,45
13	5,71	88,10	13,44
14	6,63	86,14	7,36
15	3,57	91,39	6,61
16	6,10	76,63	9,37
42	2,71	71,33	45,42
43	6,56	49,47	45,97
44	0,95	29,90	83,94
45	6,51	41,57	90,98
46	5,78	44,02	70,41
97	3,64	24,13	83,65
98	3,28	55,52	82,10
99	19,66	83,81	98,84

The DeepL results, albeit preliminary are worst especially for the stems class, of course other architectures need to be tested and optimized, however some cons of this methodology are the higher number of training images required and the much demanding computational effort in the learning phase.

Overall, these results highlight that the adopted WT + PLS-DA approach could potentially be useful to measure the ratio between leaves and stems in the basil plants controlled by the in-line RGB Vision System.

The possibility to characterize basil plant when they arrive at the production plant could give a relevant increase in the quality of the final product. To do that it is very important to have the capability to proper elaborate the RGB images acquired by the Vision System already installed in the plant. The chemometric approach gave a promising way to solve this topic.

6.2 Hyperspectral imaging (HSI)

6.2.1 Introduction

In the food industry there is an increasing need of fast and non-destructive analytical methods to evaluate the characteristics of products, especially for in-line or on-line monitoring in production plants. Hyperspectral imaging (HSI) is a powerful methodology both fast and non-destructive, as well as being possible to implement in/on-line. Moreover, from Hyperspectral images it is possible to obtain both morphological and chemical information.

In this work, a preliminary study has been done on the feasibility of applying Visible (Vis-HSI) and Near Infrared Hyperspectral imaging (NIR-HSI) for the characterization of Italian "*Pesto alla Genovese*" sauces.

Pesto samples obtained by basil coming from three different origins giving rice to pesto sauce with different characteristics, were studied. The aim was twofold: on one hand to set up a data analysis strategy to fully exploit the information carried out by HSI, and on the other one, to distinguish the different categories of pesto.

The multivariate image analysis pipeline, applied to Vis an NIR HSI, comprises a Region Of Interest (ROI) extraction from each image, a proper spectral pre-processing step, then

Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) multiset analysis was used to obtain the purest components spectral profiles and their respective relative concentration map.

Post elaboration of concentration maps of the resolved components has been applied to study their spatial distribution. In particular, a homogeneity index has been estimated for each resolved component by implementing the "homogeneity curve" method. In addition, global textural features were extracted by Gray-Level Co-Occurrence Matrix. Exploratory Principal Component Analysis (PCA) was then applied on the extracted features, allow distinguishing the different pesto categories.

As mentioned in the introduction to this chapter the smooth structure of pesto obtained by a balanced oil-water emulsion is fundamental for a good quality product, in this respect hyperspectral imaging has the potential to be used to monitor this characteristic. In fact, it joins the possibility to inspect the distribution (it can be obtained by the image of the sample) of the ingredients/phases (the chemical information come from the spectra available for each single pixel of the image). So, the use of a hyperspectral system coupled with proper data elaboration, can extract information about the chemical components and their distribution on the surface of pesto, which may be due to the different basil used in the recipe, or other processing steps. In particular, the eventual oil release can be monitored.

6.2.2 Material and Methods

6.2.2.1 Sampling

Twenty-three commercial samples of Italian sauce "*Pesto alla Genovese*" were collected directly at the production plant during the whole productive season, ranging from May to October 2021. The collected pesto samples were obtained by basil plants of three different origins called 1, 2 and 3 for confidentiality reasons. This prior information about basil was used as class label for the pesto samples, resulting in three classes into which can be potentially distinguished. The collected samples, their month of production and basil origin are reported in Table 6.2.

Table 6.2. Samples collection scheme, classes and numerosity

Month	**Number of samples**
May	3 (class 1)
June	3 (classes 1, 2 and 3)
July	5 (class 3)
August	4 (classes 2 and 3)
September	4 (class 2)
October	4 (classes 1, 2 and 3)

6.2.2.2 Instrumentation and images sampling

In the present study were used two hyperspectral cameras assembled at INRAe facilities. They cover the spectral ranges from 409 to 987 nm (24450 to 10132 cm^{-1}) and from 964 to 2494 nm (10373 to 4009 cm^{-1}), respectively (see 2.5.2). Images were acquired by pouring an aliquot of pesto onto a disposable aluminium vessel. Together with pesto as white reference a white tile was imaged to correct illumination differences, from sample to sample, and to normalize each image. Acquisition was done in reflectance mode. Then the acquired images were normalized dividing every pixel by the average signal of the white tile to compensate the eventual illumination changes between acquisitions. Subsequently images were converted in absorbance using the formula:

$$Image\ Absorbance = -\log_{10}(Image\ Reflectance) \qquad Eq.\ 6\text{-}3$$

For instrumental reasons the images in the Vis and NIR range were collected in subsequent sessions. In a preliminary session (data not reported) were tested three different possible sample presentations to the hyperspectral cameras: 1) through the glass on the bottom of the jar; 2) on

the sample contained in the jar previous removal of the first superficial layer; 3) on the sample collected and transferred to an aluminium vessel. These preliminary trials highlighted in the first case some problems of unwanted reflection on the jar glass, and in the second case some problem of unwanted shadows and non-homogeneous illumination. So, the third presentation mode was chosen to collect the images.

The Vis camera had a resolution of 1167x1600 pixels and in the range 408-987 nm, 160 spectral wavelengths were sampled. This, for each image sample a 3D array of dimensions 1167x1600x160 was obtained.

Before further elaboration a region of interest (ROI) (Figure 6-6) of dimensions 400x400 pixels from the centre of the sample was selected for each image, obtaining a 400x400x160 array, which for computational reasons, was further resized to 100x100x160.

The NIR camera had a resolution of 320x260 pixels with 256 spectral wavelengths from 964 to 2494 nm. Analogously to the Vis case, a square region of interest (ROI) of 91x91 pixels was selected, giving for each sample a data array of dimensions 91x91x256.

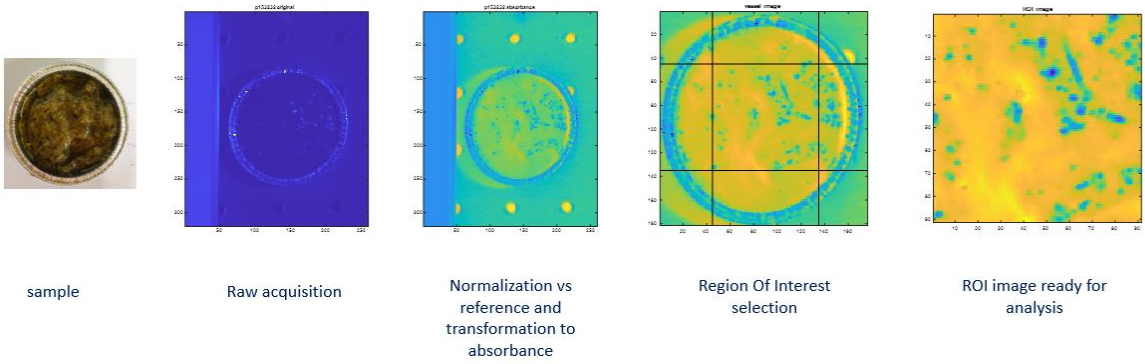


Figure 6-6 Steps to prepare image for analysis, used for both Vis and NIR cameras: from the left: original sample in aluminium box, with reference white tile, the raw image, the image normalized vs reference and transformed to absorbances, Region Of Interest selection, ROI image ready for analysis.

6.2.3 Data analysis

6.2.3.1 Spectral preprocessing

For each single image, after pixels-wise unfolding, the spectra were preprocesssed by applying SG smoothing (polynomial order 2 and window 9), baseline correction (Whittaker lambda=100 sigma 0.001) and normalization by Multiplicative Scatter Correction (MSC, see ref. 13 in chapter bibliography) using as reference spectrum the average one.

6.2.3.1 MCR-ALS decomposition

The use of MCR-ALS in image analysis has been introduced in paragraph 3.6.1. In this case, for computational constraints due to the image dimensions, the MCR-ALS model was calculated on six representative images, i.e. corresponding to samples number 2, 4, 5, 8, 12 and 16, by using the multiset modality (Figure 6-7 top). The Vis images were not pre-processed and a three components MCR model was calculated (applying non-negativity constrains on both **C** and **S** matrices).

For the NIR images to better resolve the spectral profile of purest components we proceeded as follow:

i) as a first step a single pre-processed unfolded image was decomposed in principal components (by singular value decomposition), then the pixels carrying essential information [9] were individuated by applying the convex-hull in the normalized (dividing

by the first component to make the PC space convex) scores space (component 2 vs. component 3). Eight pixels belonged to the convex-hull;

ii) The spectra corresponding to the eight pixels were decomposed by MCR-ALS imposing non-negativity constrains on both **C** and **S** matrices, retaining three components;

iii) The **S**opt matrix, obtained in step ii) was then used as spectral initial guess in the MCR-ALS of the multiset composed by concatenation row-wise of the six unfolded and pre-processed representative images. In addition to non-negativity, for one of the three component a selectivity constrain was also imposed (trying to recover an aqueous phase component, we imposed zero values in spectral wavelengths where water does not absorb). In this way new **C**opt and **S**opt matrices were obtained;

iv) The **C** matrices for all the other images (the other samples) have been calculated (after applying unfolding and preprocessing) by inverting the MCR equation and by using the **S**opt obtained in iii) (Figure 6-7 bottom).

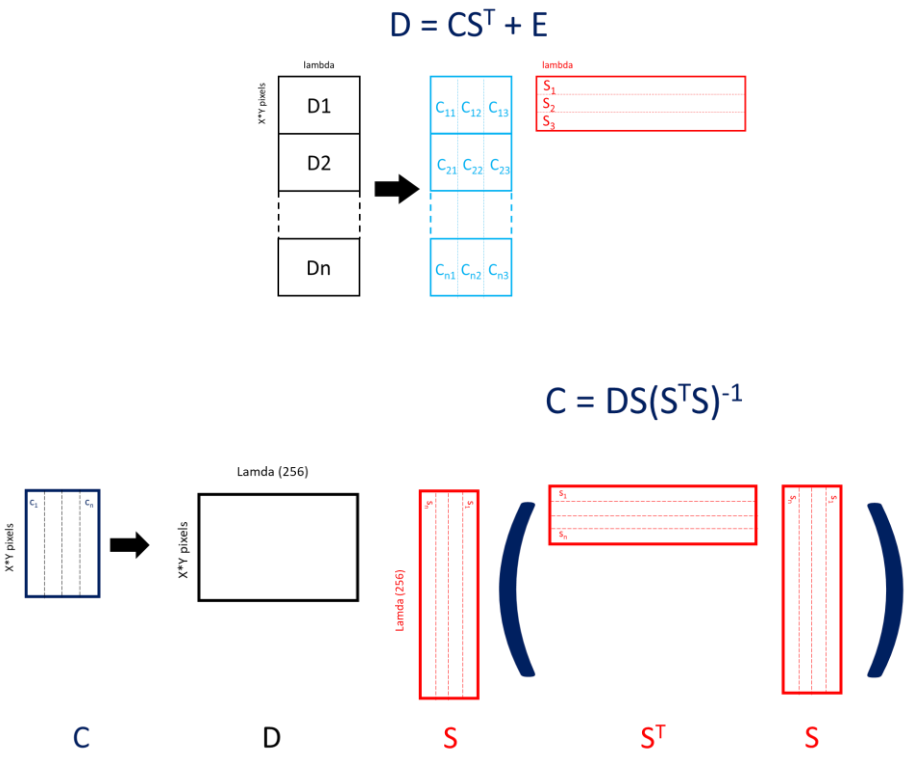


Figure 6-7. Scheme of application of MCR-ALS to the multiset (top) composed by the images of six representative samples. Bottom) obtaining the concentration matrix **C** for the remaining samples by MCR model inversion.

The images of the refolded concentration matrix (concentration maps) of the purest components, for both Vis and NIR MCR-ALS models, and their relative spectra are reported in Figure 6-8.

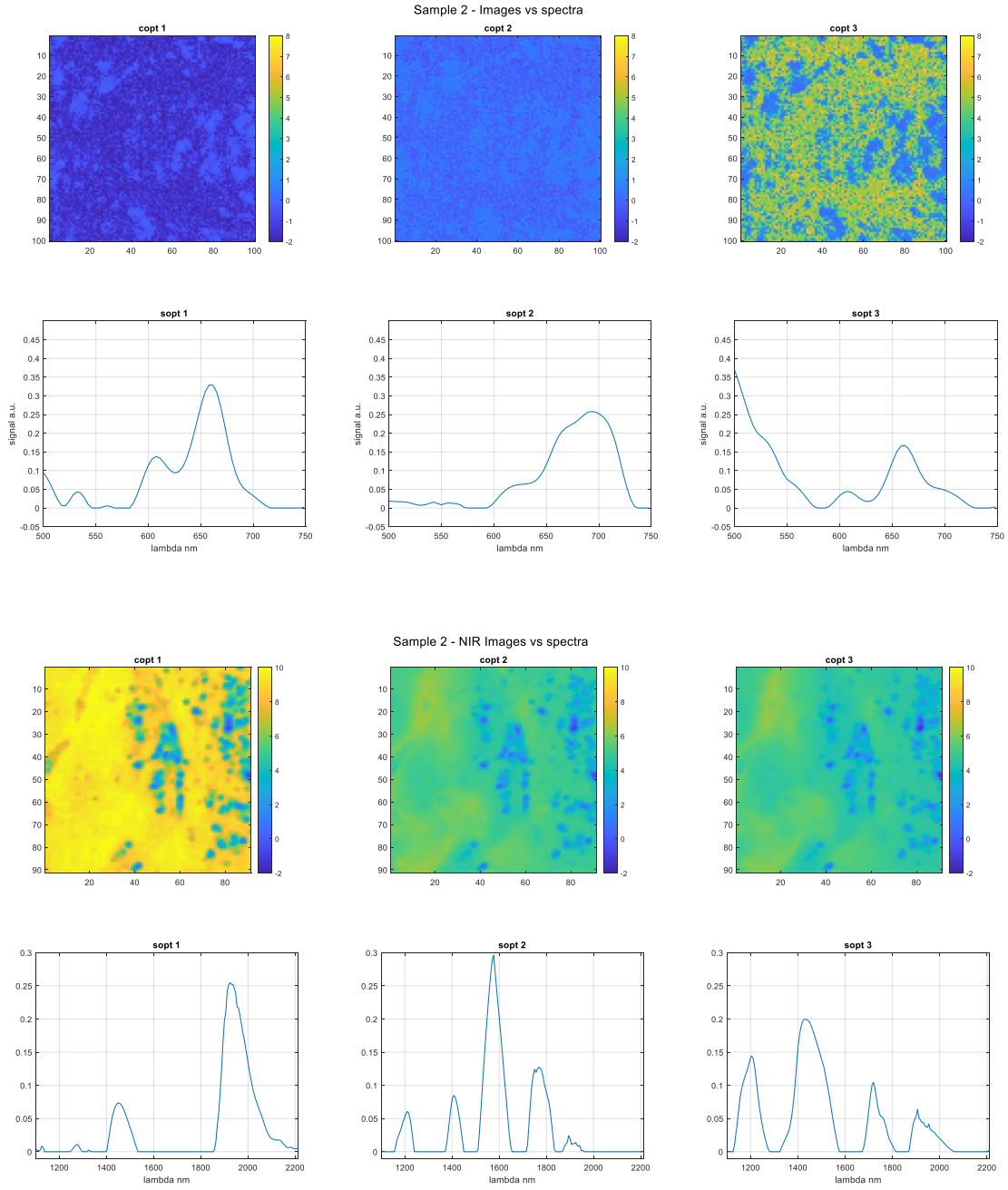


Figure 6-8. Images of the purest resolved components and their correspondent spectra for sample 2, one of the samples used to build the MRC-ALS model. In the upper image the Vis model (concentration map and resolved spectra) and in the lower image the NIR model (concentration map and resolved spectra). In both cases, the components are reported in order from one to three respectively from left to right.

In the Vis the spectral profile of the first pure component shows two bands that can be ascribed to the absorption of the chlorophylls, respectively “chlorophyll a” at around 662 nm, and “chlorophyll b” at 642 nm. The spectral profiles of the second and the third components could be attributed to the absorption of oil [10]. The distribution is rather homogeneous and similar for component 1 and 2, while specular for component 3 (i.e. where components 1 and 2 show the higher intensity component 3 show the lower).

In the NIR, the spectral profiles of the three purest components remind to absorption bands of water, proteins, and oil (as described in the method chapter 2). In details, the first component presents the two typical absorption bands of water, the second component present the most intense band in the spectral region where proteins absorb, and some less intense bands in the spectral regions where water and lipids absorb. This could suggest that the second component may be representative of the emulsion phase. The third component shows bands mainly in the spectral regions where lipids absorb (thus could represent oil). The spatial distribution of the three components is rather similar, but there is a region in which aqueous phase prevails (intense yellow colour in concentration map of component one) and the blue spots represent regions where all components have low concentration values.

6.2.3.1 Post Processing of concentration maps

To characterize and differentiate the pesto samples disposing of global features for each sample is useful. To this aim, some features were calculated by the concentration maps of the purest components treating them by image analysis tools. In particular, the two approaches described in chapter 3 were applied on the six concentration maps for each sample obtained by MCR-ALS of NIR and Vis imaging data, respectively (three for each).

The first approach was the calculation of image features by using the Haralick method on the Gray-Level Co-Occurrence Matrix (GLCM) [11)]. The features were calculated as detailed in Table 3-1 of chapter 3 exploring different pixel neighbours' distances, namely 1, 2, 4 and 8, and different grey levels, such as 8, 16, 32 and 64. At the end, merging Vis and NIR data, a matrix of 23 samples x 768 columns (features) was obtained and was used for further PCA explorative analysis.

The second approach was the Homogeneity calculation by Continuous – Level Moving Block (CLMB) method by using the methodology proposed in [12], which is based on the Macropixel analysis methodology, already presented in chapter 3.

The formula was applied to all the 6 images i.e. concentrations maps of the purest components of each sample, giving as result a matrix holding the homogeneity index in percentage of dimensions of 23x6. On this matrix an explorative PCA was done.

6.2.4 Results and discussion

The spectra of the pure components decomposed by MCR gave some chemical information on the images. In fact, observing the spectra in the visible range (Figure 6-9) was possible to note that component 1 cover the absorbance of chlorophylls a and b, relative to basil, while component 2 and 3 could be relative to the olive oil pigments (including chlorophylls).

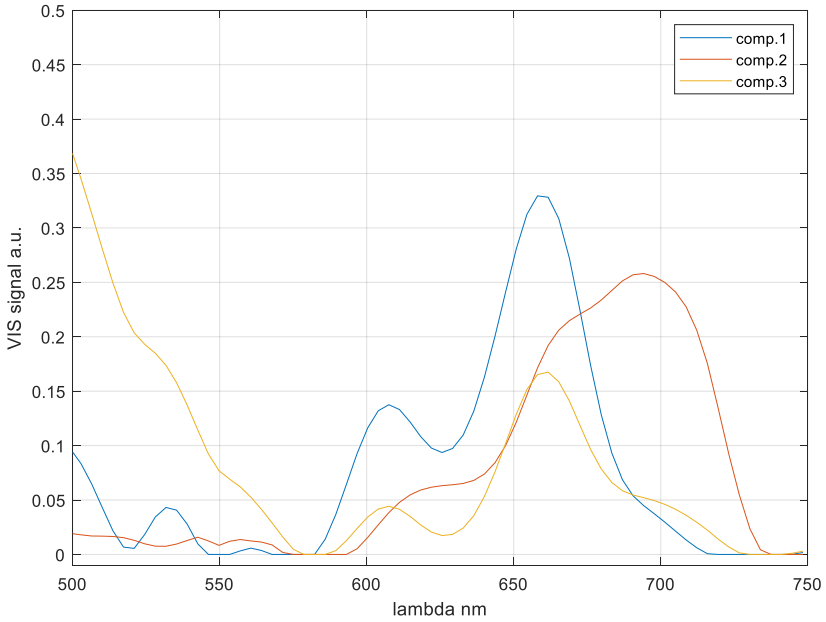


Figure 6-9. Spectra of pure components in the VIS hyperspectral images decomposed by MCR-ALS.

Observing the results obtained for the NIR range, the three pure components (Figure 6-10) could be ascribed to water (absorption bands at about 1450 and 1940 nm), lipids (absorption bands at about 1200, 1700 and 2300 nm) and proteins (absorption bands from 2050 to 2180 nm).

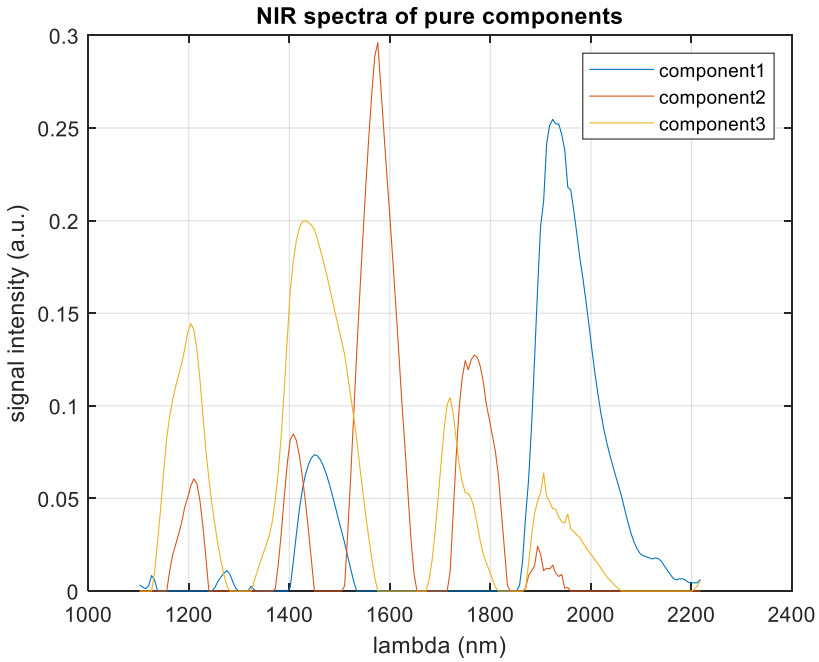


Figure 6-10. Spectra of pure components in the NIR hyperspectral images decomposed by MCR-ALS.

The respective concentration maps, obtained for each sample are shown on Figure 6-9 and Figure 6-10. The general observations drawn for the six samples on which the MCR models were built (see 6.2.3.2) still holds. In addition, it can be seen some differences from sample to sample with respect to the spatial distribution of one or more components, e.g. a different degree of homogeneity for component 3 of NIR (attributed to lipids).

The features calculated on these concentration maps by the two approaches described in the chapter 3 were used for two distinct explorative PCAs.

The PCA on the Haralick features (Figure 6-11) shows in both PC1 *vs* PC2 and PC3 *vs* PC4 scores plots a poor separation between the three classes. The highest PCs have been also inspected but not reported because did not add more information. Samples number 18 and 23 appear to be very different from the other samples. For sample 18 in fact, looking at its concentration maps, a different distribution for components 1 and 2 could be observed, especially for the Vis (pointing to different distribution of chlorophyll pigments and eventually in segregation of basil plant residues) and to a less extent for the NIR components as well, with respect to the other samples. For sample 23 differences with respect to other samples are more evident in NIR images, especially for components 1 and 3, related to chlorophylls and oil.

The reason explaining this behaviour need further investigation.

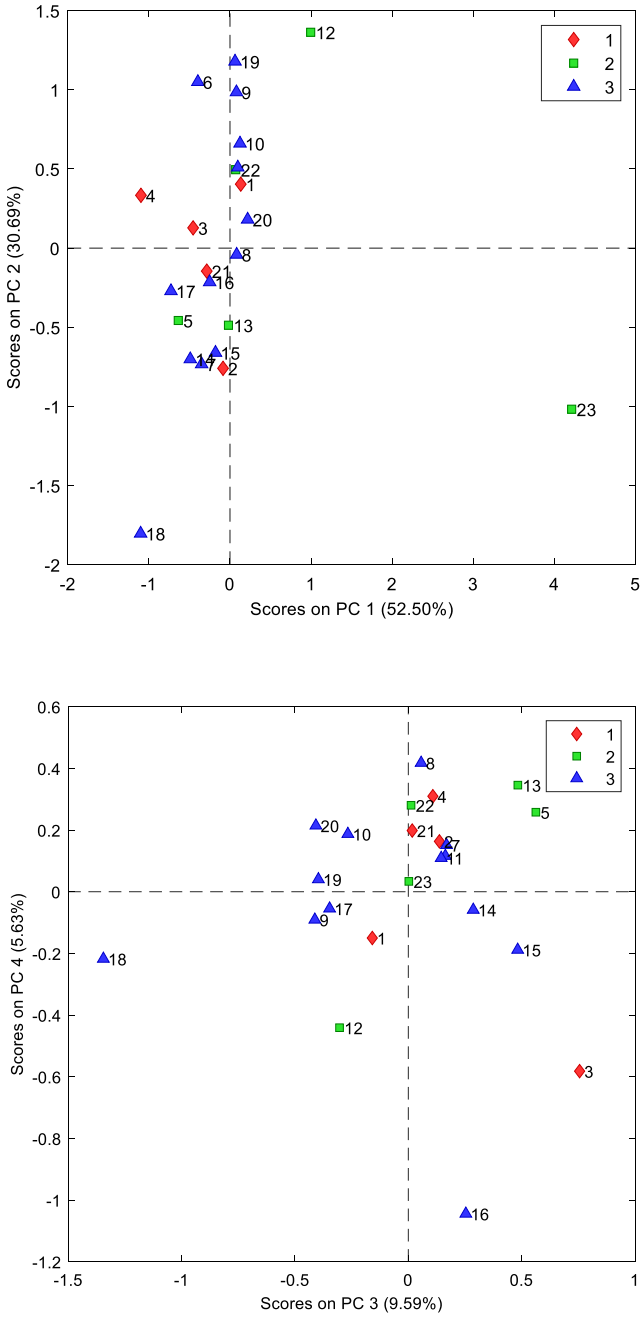
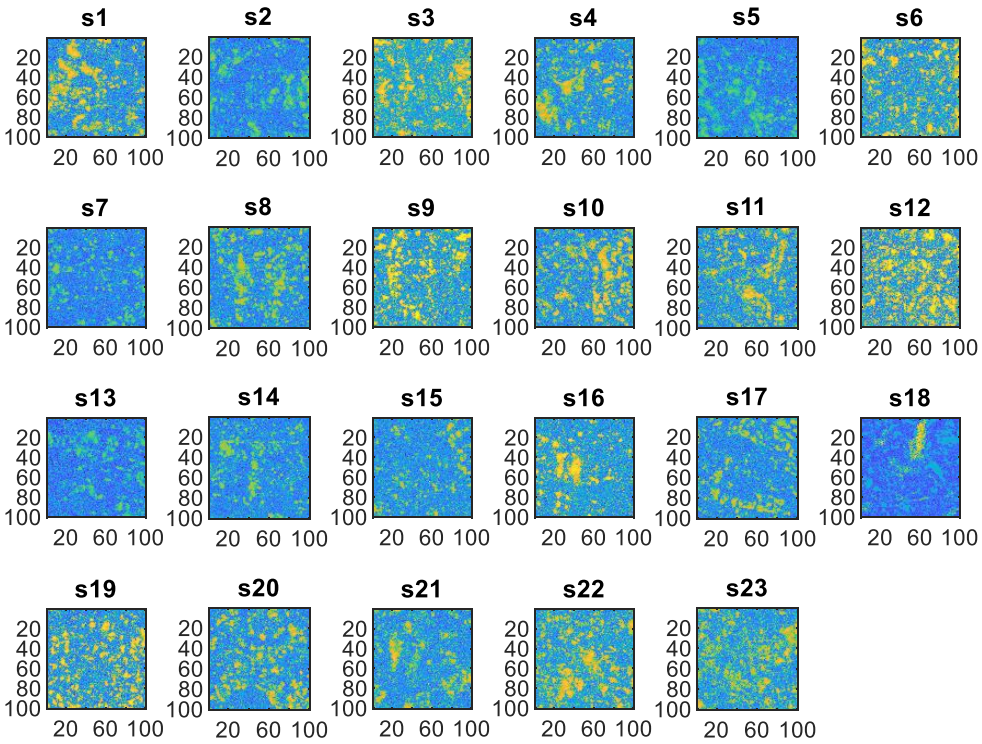
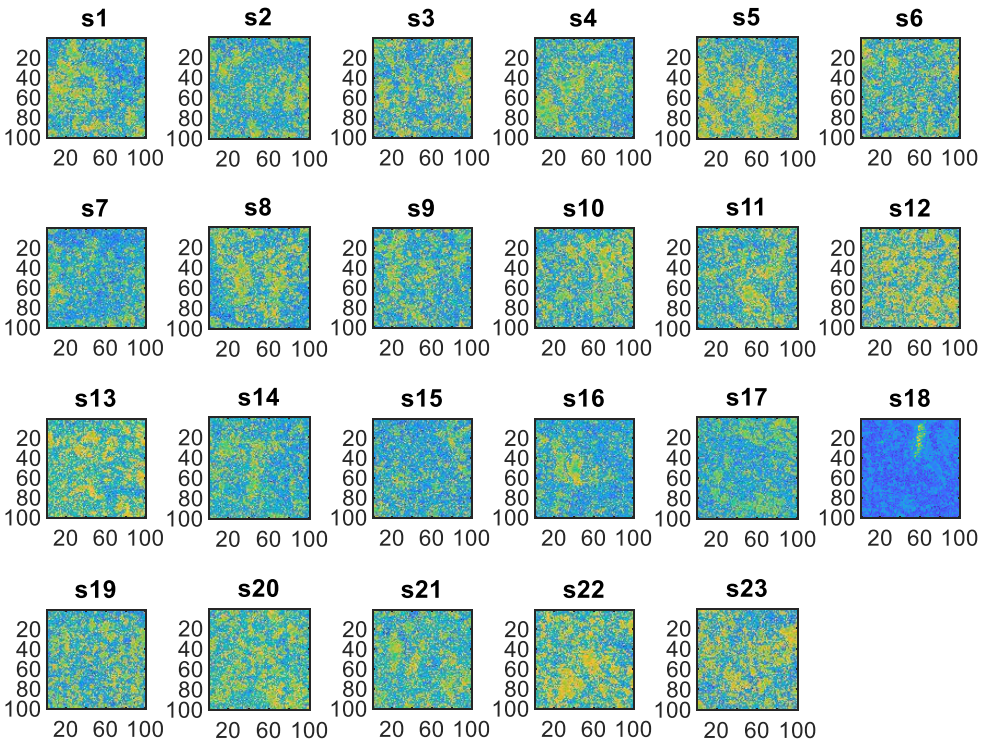


Figure 6-11. Haralick's classical features calculated on both Vis and NIR images of pure components. Score plot of PC1 and PC2 (top) and PC3 and PC4 (bottom) of explorative PCA. The different colours indicate the three pesto classes related to the basil origin.


VIS - All samples - Component 1
s1
s2
s3
s4
s5
s6
s7
s8
s9
s10
s11
s12
s13
s14
s15
s16
s17
s18
s19
s20
s21
s22
s23
20 40 60 80 100
20 60 100


VIS - All samples - Component 2
s1
s2
s3
s4
s5
s6
s7
s8
s9
s10
s11
s12
s13
s14
s15
s16
s17
s18
s19
s20
s21
s22
s23
20 40 60 80 100
20 60 100

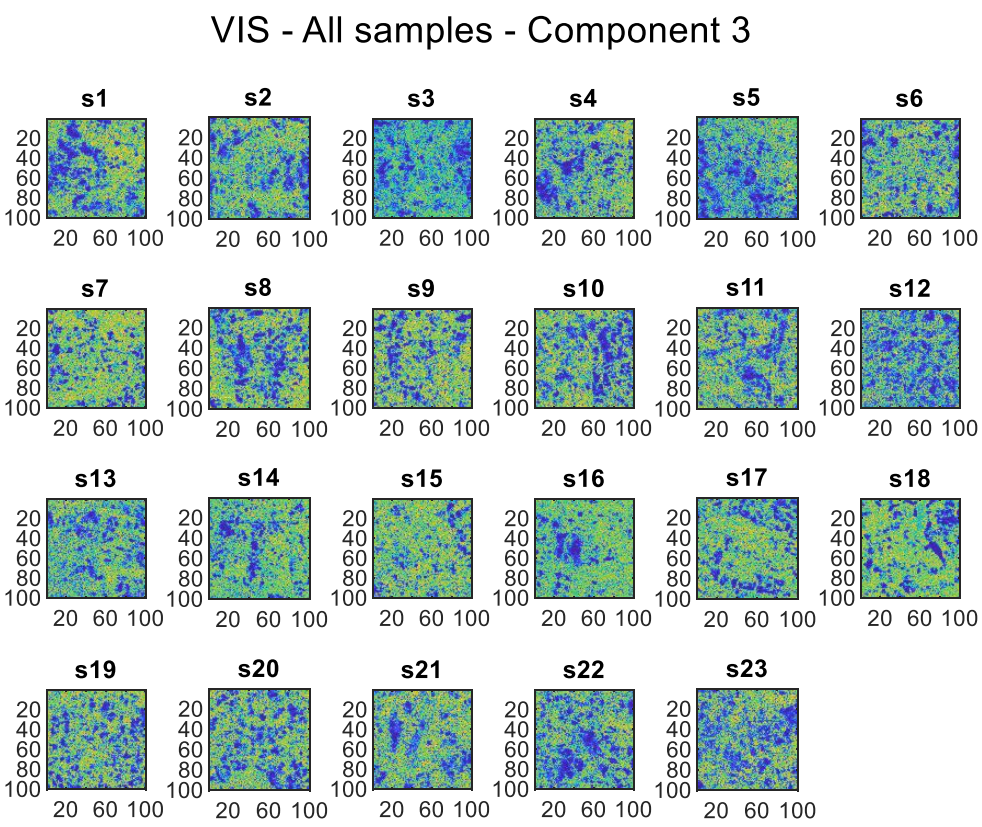
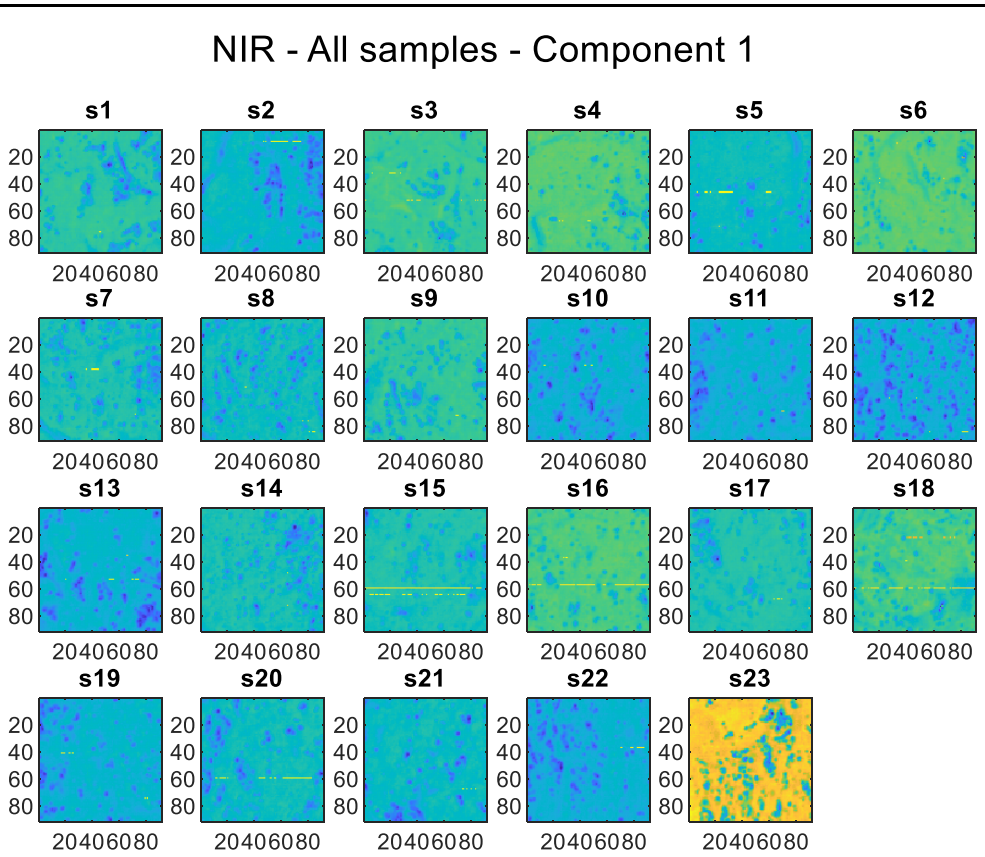
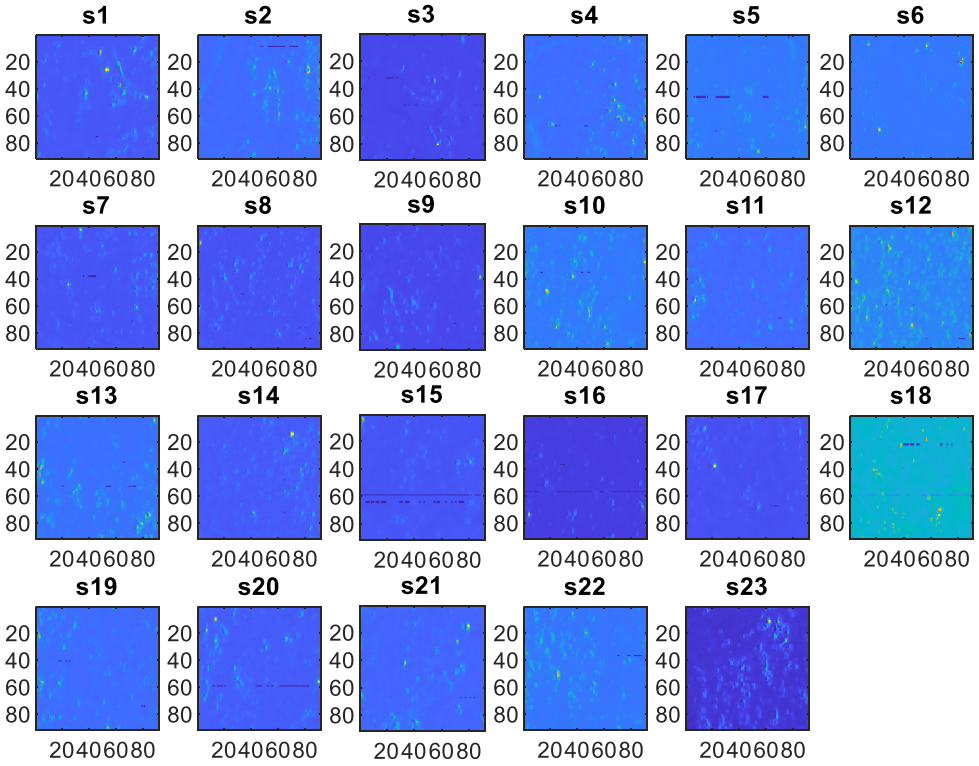


Figure 6-12. Reconstructed images of MCR-ALS pure components of all VIS hyperspectral images of all samples: respectively from top to down the three components.


NIR - All samples - Component 1
s1
s2
s3
s4
s5
s6
s7
s8
s9
s10
s11
s12
s13
s14
s15
s16
s17
s18
s19
s20
s21
s22
s23
20
40
60
80
20406080


NIR - All samples - Component 2
s1
s2
s3
s4
s5
s6
s7
s8
s9
s10
s11
s12
s13
s14
s15
s16
s17
s18
s19
s20
s21
s22
s23
20
40
60
80
20406080

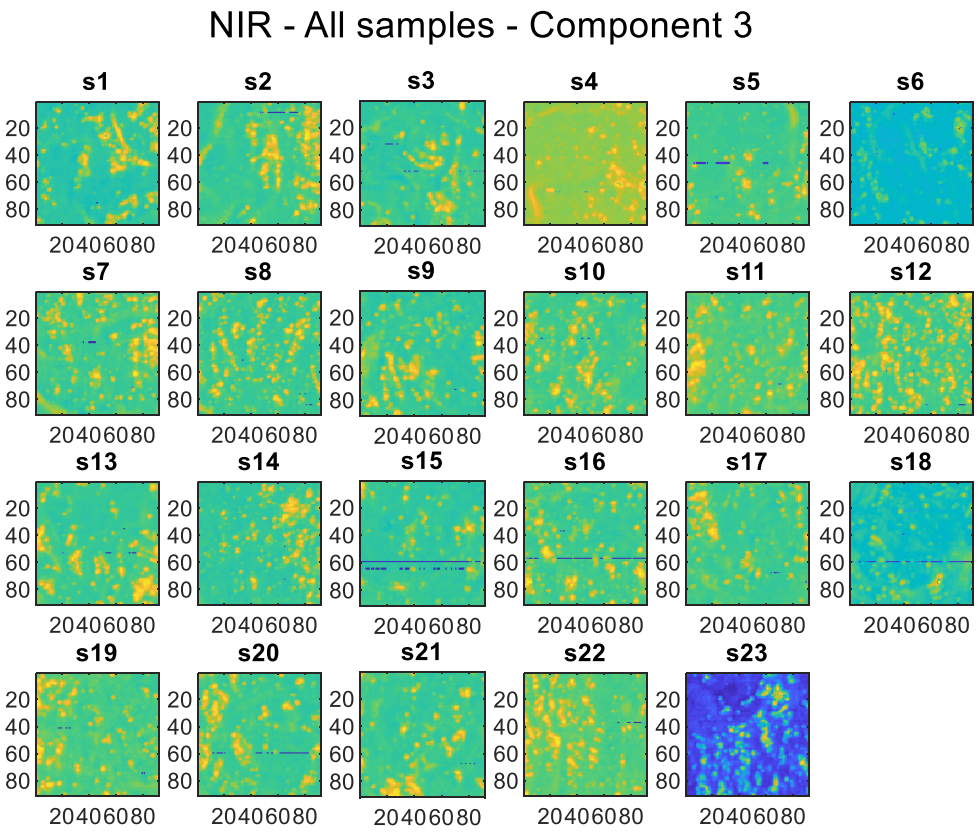


Figure 6-13. Reconstructed images of MCR-ALS pure components of all NIR hyperspectral images of all samples.

The explorative PCA calculated on the homogeneity index data (Figure 6-14), showed more overlap among different pesto classes; only class 1 samples (except one) are localized at most negative PC1 values with respect to the other two classes.

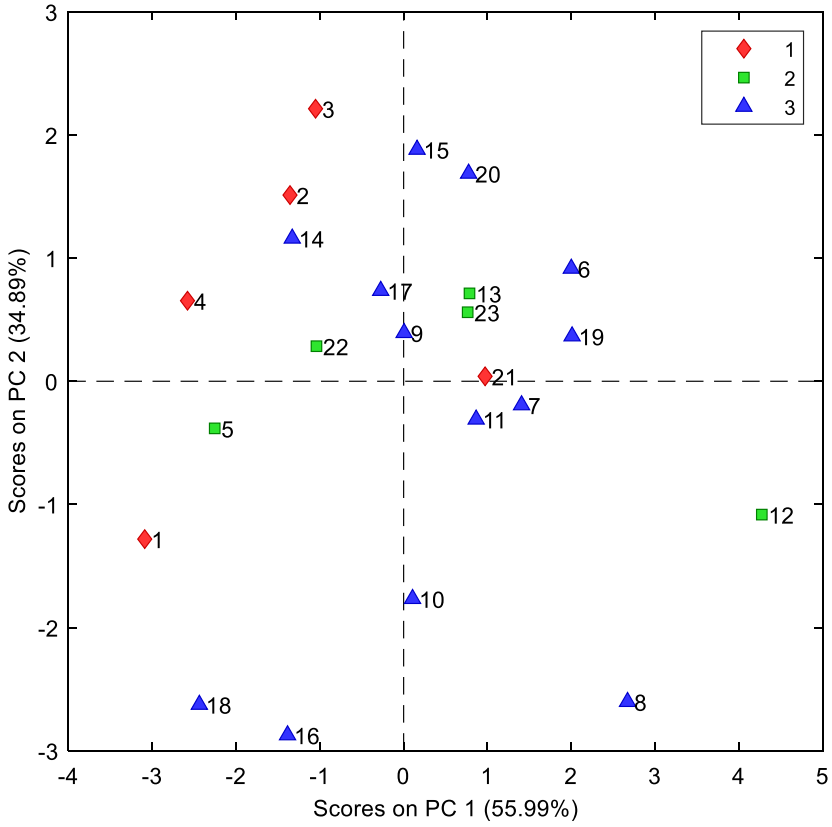


Figure 6-14. Score plot of explorative PCA done on homogeneity feature data of VIS and NIR images. The three colours indicate the pesto classes.

6.3 Conclusions

Hyperspectral imaging (HSI) is a very powerful tool to collect simultaneously morphological and chemical information of a sample. Its use is increasingly spreading due to the camera's affordability in terms of costs and performances and the augmented computing power and methodologies.

In the case of pesto characterisation their very complex matrix was a challenging topic to solve. The use of more sophisticated elaboration technique like MCR-ALS helped to extract information from the images giving as result images of pure components, which were putatively attributed based on their resolved spectra. Inspection of this distribution maps, sample by sample, provide a depth insight on how smooth the structure of pesto is. However, the global features extraction, either by the Haralick method or by the homogeneity index, did not brin to a clear separation in the three pesto classes. It could be that the differences in basil origin is not so crucial for the final pesto structure, and the pesto processing is successful in providing a stable product. On the other hand, some differences emerged for some few peculiar samples, such as number 18 and 23 where highlighted.

1 Salager, J. L., Loaiza-Maldonado, I., Minana-Perez, M., & Silva, F. (1982). Surfactant-oil-water systems near the affinity inversion part I: relationship between equilibrium phase

behaviour and emulsion type and stability. JOURNAL OF DISPERSION SCIENCE ANDTECHNOLOGY, 3(3), 279-292.

2 Khan MJ, Khan H S, Yousaf A, Khurshid K, and Abbas A. Modern trends in hyperspectral image analysis: A review. 2018; 6: 14118-14129.

3 Otsu N. A threshold selection method from grey-level histograms.IEEE Transactions on Systems, Man, and Cybernetics. 1979; 9(1): 62–66 doi: 10.1109/TSMC.1979.4310076. S2CID 15326934

4 https://www.sensure.it/en/azienda (January 2024)

5 Liu D, Sun DW, Zeng XA Recent advances in wavelength selection techniques for hyperspectral image processing in the food industry. Food and Bioprocess Technology. 2014; 7(2): 307-323. doi: 10.1007/s11947-013-1193-6

6 Fowler JE. The redundant discrete wavelet transform and additive noise., in IEEE Signal Processing Letters 2005;12(9): 629-632. oi: 10.1109/LSP.2005.853048.

7 Akansu AN. and Liu Y. On Signal Decomposition Techniques, Optical Engineering. 1991;12-920

8 Shensa MJ. The Discrete Wavelet Transform: Wedding the A Trous and Mallat Algorithms, IEEE Transactions on Signal Processing.1992; 40(10):

9 Ruckebusch C, Vitale R, Ghaffari M, Hugelier S, and Omidikia N. Perspective on essential information in multivariate curve resolution. TrAC Trends in Analytical Chemistry. 2020; 132: 116044.

10 Philippidis A Poulakis E, Papadaki A and Velegrakis M. Comparative Study using Raman and Visible Spectroscopy of Cretan Extra Virgin Olive Oil Adulteration with Sunflower Oil, Analytical Letters. 2017; 50(7): 1182-1195.doi: 10.1080/00032719.2016.1208212

11 Haralick RM, Shanmugam K, and Dinstein I. Textural Features for Image Classification, IEEE Trans. on Systems, Man, and Cybernetics. 1973; SMC-3(6):610-621

12 de Moura França et al. "Evaluation and assessment of homogeneity in images. Part 1: Unique homogeneity percentage for binary images". Chemometrics and Intelligent Laboratory Systems. 2017; 171: 26–39.

7A FEASIBILITY STUDY TOWARDS THE ON-LINE QUALITY ASSESSMENT OF PESTO SAUCE PRODUCTION BY NIR AND CHEMOMETRICS

This work has been done in collaboration with Daniele Tanzilli, a PhD student of Professor Marina Cocchi and it is the object of Publication n°3 (reported in Chapter 8), to which the reader is referred for more details. I was responsible of data acquisition, curation, exploratory analysis. Equal contribution was given to results presentation, validation and discussion, writing, editing. Daniele, was responsible of methods development in particular for data synchronization, preprocessing and predictive modelling.

In the following, I will report just a part of the whole work done, such as the preliminary feasibility analysis (which is not present in the published paper) and the multivariate control charts, while for the on-line prediction only a Table summarizing the obtained models (including some which are not present in the publication) is shown.

7.1 Introduction

The texture of pesto is a delicate equilibrium of an emulsion of oil in water, protein matrix, pieces of basil (leaves and stems) and of cashew nuts [1]. Its stability depends on proper raw materials characteristics and proportion, and preparation process (cutting, mixing and thermal treatment). The control of the process in its crucial steps became so an important part of the production, from one hand to maintain the designed quality, to the other hand to detect potential critical conditions with the aim of minimize wastes or production stops. The process control could be achieved with on-line monitoring systems and models, and NIR spectroscopy is widely used in food processes with appropriate chemometric models developed.

A preliminary part of this part of the study was done to evaluate the feasibility of NIRS to gather compositional/structural information which will then allow, aided by chemometrics modelling, predicting the pesto sauce characteristics, and particularly its emulsion stability. In fact, the reference method used in R&D Lab to measure of the emulsion stability of the pesto structure, during the development of new recipes or technologies, is based on analysis with a dedicated type of centrifuge (LUMiSizer®, see Chapter 2). This centrifuge is able, during the centrifugation process, to measure the speed and the amount of the pesto phases separation thus allowing to determine the emulsion stability. However, it is not appropriate in a production context because each measure requires several hours and trained personnel. So, the need to evaluate a more rapid and “easy to use” technique like NIRS. The evaluation study was done in Lab with an off-

line bench instrument, but a future step could be to exploit the methodology in the plant by using on-line NIR instrument.

In the published work, indeed a feasibility study was done to establish how an on-line NIRS system (already installed in pesto production plant) could help to real-time monitor the intermediate/final product quality during production by using Multivariate Statistical Process Control (MSPC) methodology. To this aim, data from one pesto production campaign was analysed by applying both multivariate control chart (MCC) based on Principal Component Analysis (PCA), and PLS regression-based models to calibrate specific properties of finished product, i.e. pesto. Since NIR spectra are collected on-line, once models are developed, they can be applied real-time in prediction and monitoring for early estimation of product quality and for early detection of any departure from normal operating condition during processing.

7.2 Materials ad methods

7.2.1 NIR feasibility study at R&D Lab scale

During the 2022 pesto production campaign 182 samples of "*Pesto alla Genovese*" were collected over three months, just after the production, in Rubbiano plant. Their emulsion stability, expressed as instability index, was measured by the reference method in the R&D Lab in Parma.

Samples were collected in three distinct production phases: 1) at the start of a production lot; 2) in the middle of a production lot; and 3) at the re-start after a production stop due to some issues and maintenance operations. In this last case, part of the product lasted in a tank for some time (variable duration depending on the type of issue to be solved) at higher temperature under continuous mixing, conditions that could affect the stability of the emulsion in the final product.

To be representative, for each sample eight NIR spectra (in eight distinct positions by rotating the sampling cup) were recorded in the range 400-2500 nm, by the benchtop instrument. The average spectrum of the eight acquired spectra was used to assemble the spectral data set.

The spectral dataset was divided in two blocks, by splitting the wavelengths regions, related to the two detectors of the NIR instrument. Respectively from 400 to 1100 nm the silicon detector and from 1100.5 to 2500 nm the lead sulphide detector (Figure 7.1). The part of the spectra from 2360,5 to 2500 nm was removed because it is affected by high noise.

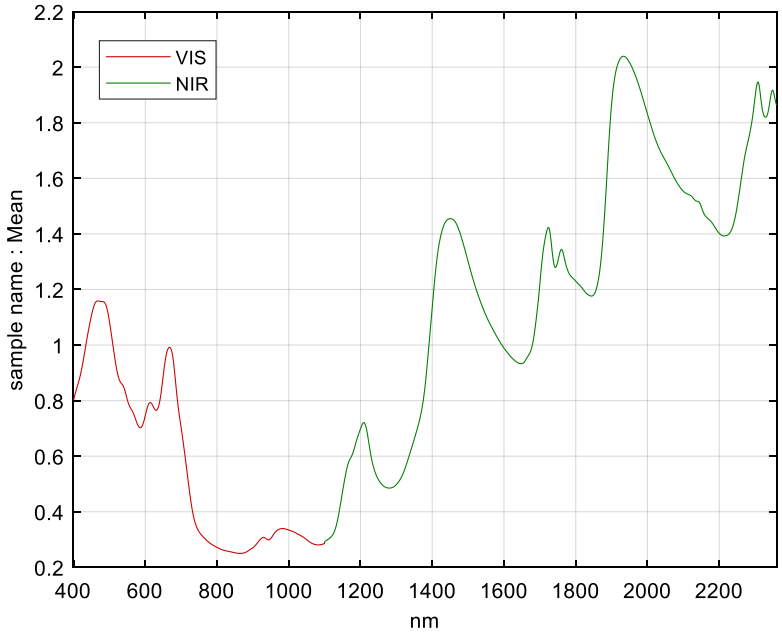


Figure 7.1. VIS-NIR spectrum example with the colours that indicates the separation into two blocks based on the two instrument detectors (red silicon detector, green lead sulphide detector).

Several spectral preprocessing was tested on both blocks (i.e. SNV and Savitsky-Golay first and second derivatives and their combinations) and at the end the first derivative (2nd polynomial, 15 points) was selected, for both of them, showing a better distinction in PCA space of the three production phases.

Samples were then divided in a calibration set (127 samples) and a validation set (55 samples) by using the DUPLEX algorithm.

Multiblock PCA was applied after block-scaling to unit variance (including mean centring).

7.2.2 Monitoring of semifinished product by on-line NIR

Semifinished pesto is a mix of oil, salt and cut basil, produced in the first part of the production process before the addition of the other ingredients. The NIR on-line probe recorded the spectra of semifinished product in a spectral range from 1100 to 1650 nm.

NIR spectra were pre-processed to remove effects, such as scattering, introducing variability not linked with information to be retrieved, and/or to enhance extractable information. In particular, Savitzky–Golay 2nd derivative and mean centering were applied prior to exploratory Principal Component Analysis and multivariate control charts building.

The dataset had been split in calibration and test sets manually, considering Normal Operative Conditions (NOC) observations, subdividing each period without production stops, as follows: the first part (about 65%) consisted of temporally contiguous points in the calibration set; and the second part (about 35%) was in the test set. In this way, we mimicked the real situation of continuous monitoring where samples to be predicted came after in time for each period. Observations not in NOC, as highlighted by exploratory PCA, were all included in the test set.

To estimate the correct number of PCs, cross-validation was performed with a venetian blind scheme with ten splits. The MSPC charts were based on two parameters: Hotelling T^2, which described the distance of a sample in the model space, and Q, which defined the distance of a sample from the model space. In other words, if a sample had high T^2 values, the model was able to describe it, but the distance between the sample and the centre of the model was high, i.e., it showed an extreme behaviour. On the other hand, if a sample was characterized by high Q values, the model was not able to describe the sample properly, hence the correlation structure of variables was different from the other samples. To assess if a sample was extreme or anomalous, signifying a departure from normal operative conditions for both control charts, the acceptance limits had to be estimated. The T^2 limit was obtained based on Hotelling's T^2

distribution, whereas the Q limit was based on χ^2 distribution and was calculated either with Jackson and Mudholkar approximation or the Box method [2, 3]

7.3 Results and discussions

7.3.1 Results of NIR feasibility study

The graph of the instability index (Figure 7.2) indicates that most of the un-stable samples (higher instability index) refer to samples collected after a stop of the production line (represented by blue triangles in Figure 7.2). The samples at the production start (shown as red diamonds) had all an extremely low instability index during all the production periods, thus good structure of the emulsion. The samples collected in the middle of the production period (green squares) showed just few unstable samples.

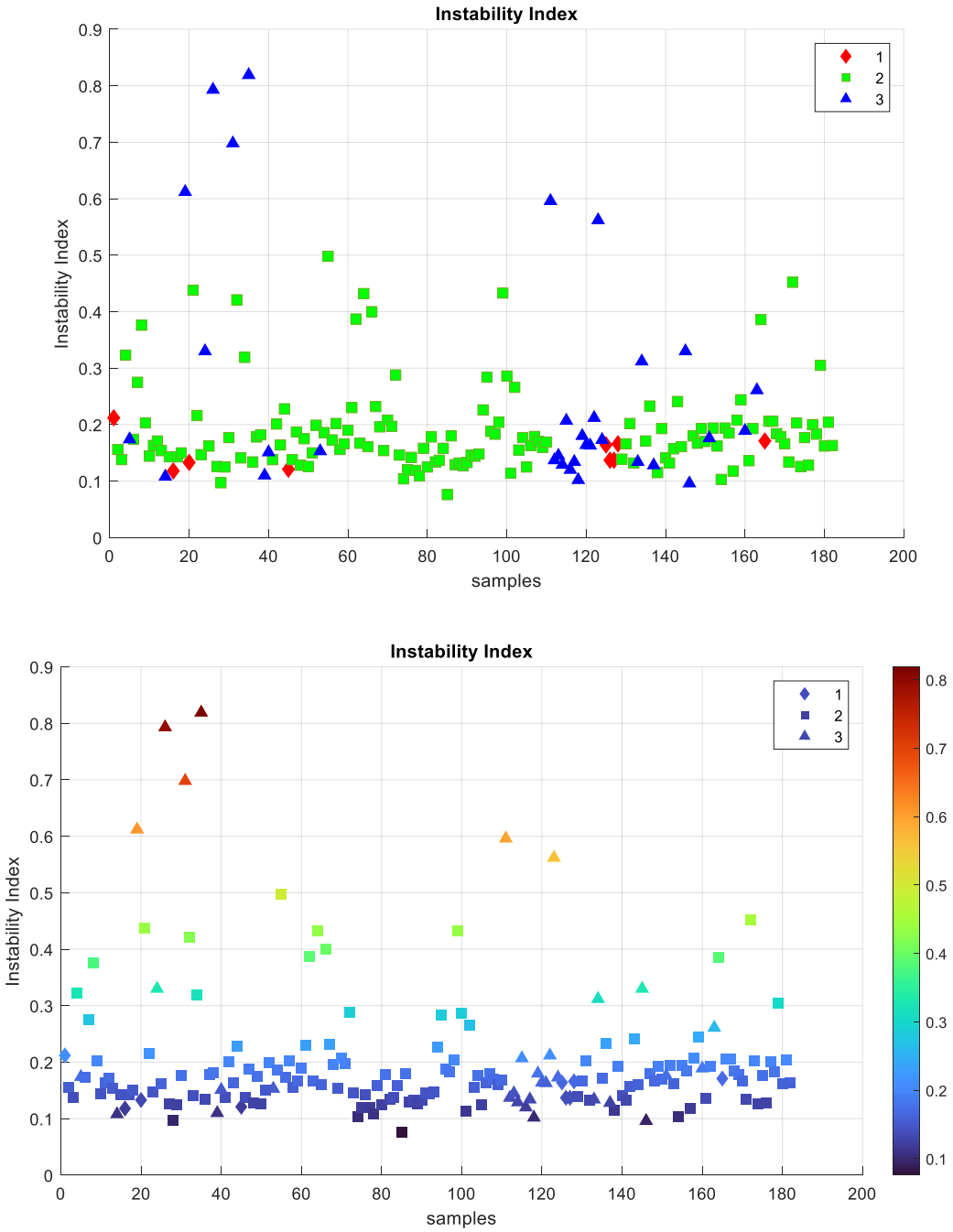


Figure 7.2. Instability index of all the samples analysed. Top plot (a): samples have different symbols and colour depending on the production phases: production start) red diamonds; middle of production) green squares; after production stops) blue triangles. Bottom plot (b): samples are coloured by instability index values (see colour bar) and different symbols are used basing on production phases as in (a) with diamond, squares and triangles indicating start, middle and after stops, respectively.

Observing the score plot of PC1 vs PC2 of the explorative PCA done on NIR spectra (Figure 7.3) is possible to note that part of the samples collected after a stop and re-start in production (blue triangles) had very negative values of PC1 and are so clearly separated from the others. These samples had higher values of instability index (Figure 7.3). On the other hand, most of the samples collected at the start (red diamonds) and in the middle of production (green squares) had positive or slightly negative values of PC1. An interpretation on the differences between these samples could be made evaluating the spectral wavelengths in the loading plot of PC1 and PC2 (Figure 7.4). Samples with negative values of PC1, so in the left part of the score plot, had lower levels of chlorophyll (band at 730 nm), higher level of water (spectral region around 1400 nm and

2000 nm where water absorption take place) and lower level of lipids (bands around 1900 nm and 2300 nm). The differences observed in PC2 was more related to the chlorophyll and proteins content.

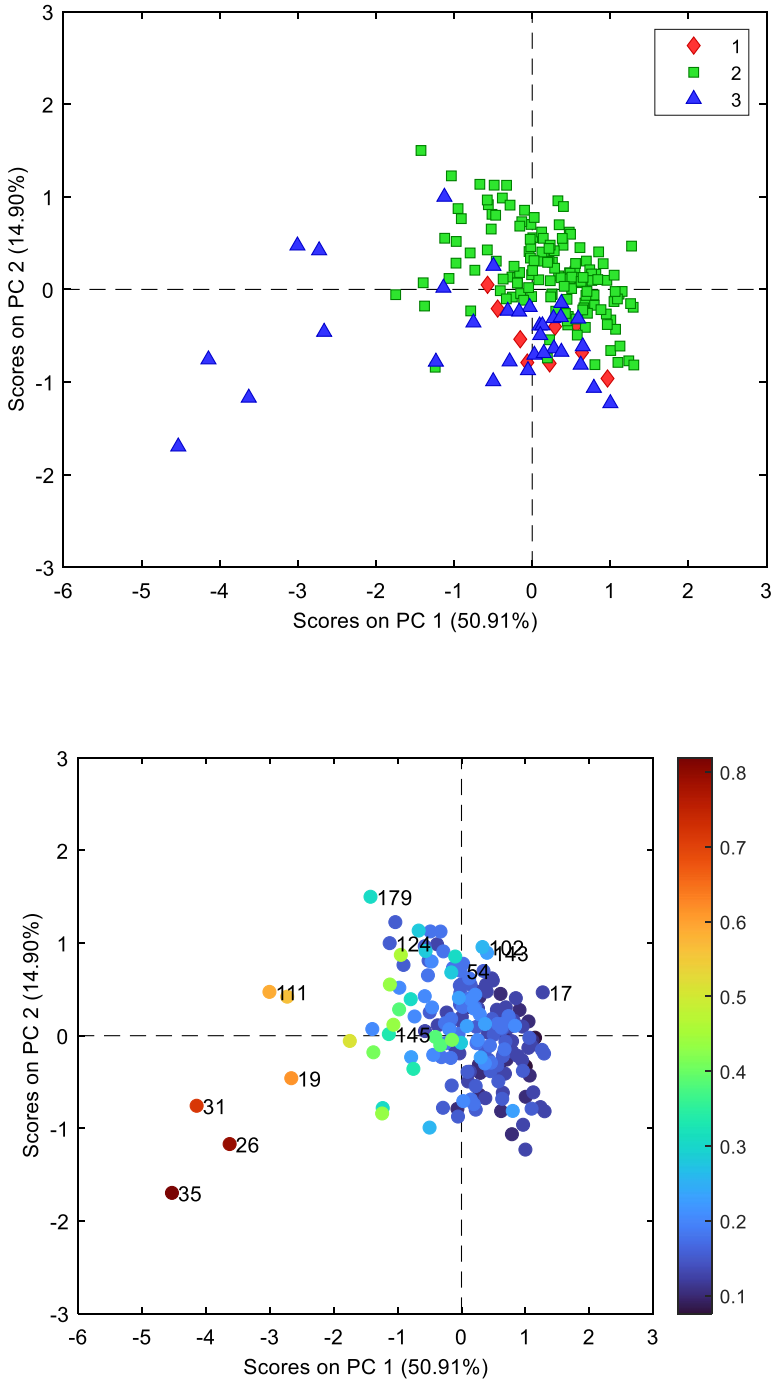


Figure 7.3. Explorative PCA of VIS-NIR spectra. Score plot of PC1 vs PC2. Red diamonds (middle of production), green squares (start of production) and blue triangles (restart after production stop) represent respectively samples collected at start, middle, and re-start of production.

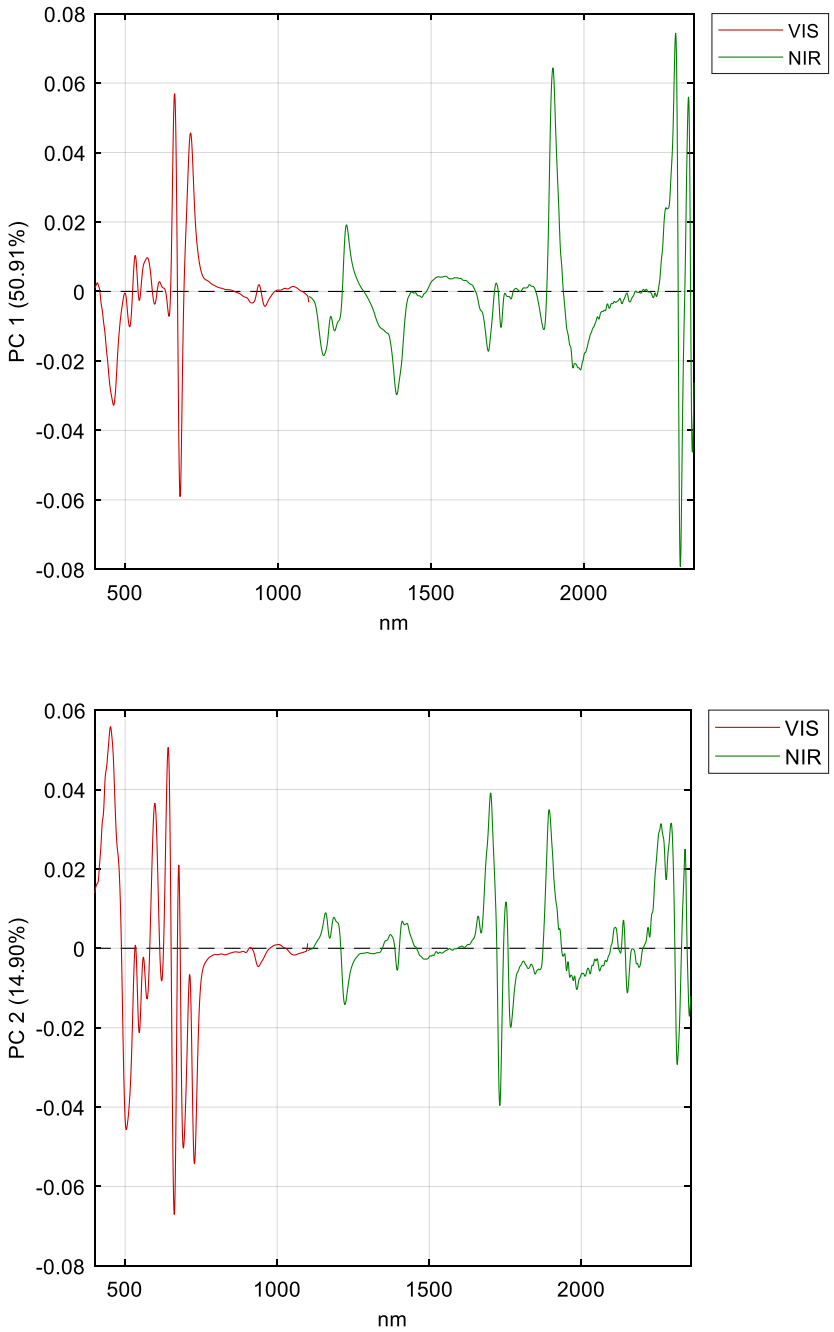


Figure 7.4. Loading plot of PC1 (top) and PC2 (bottom).

The Hoteling T^2 (Figure 7.5) plot, which show how extreme are samples in the overall PCA model, was also inspected As it can be observed in the figure, the samples having higher instability index fall over the T^2 critical limit and are so indicated as very extreme samples and most of them were collected after productions stops/re-starts. Most of these samples were the ones located at negative PC1 vs PC2 PCA scores plot (far from the others, e.g. 26, 31 and 35.

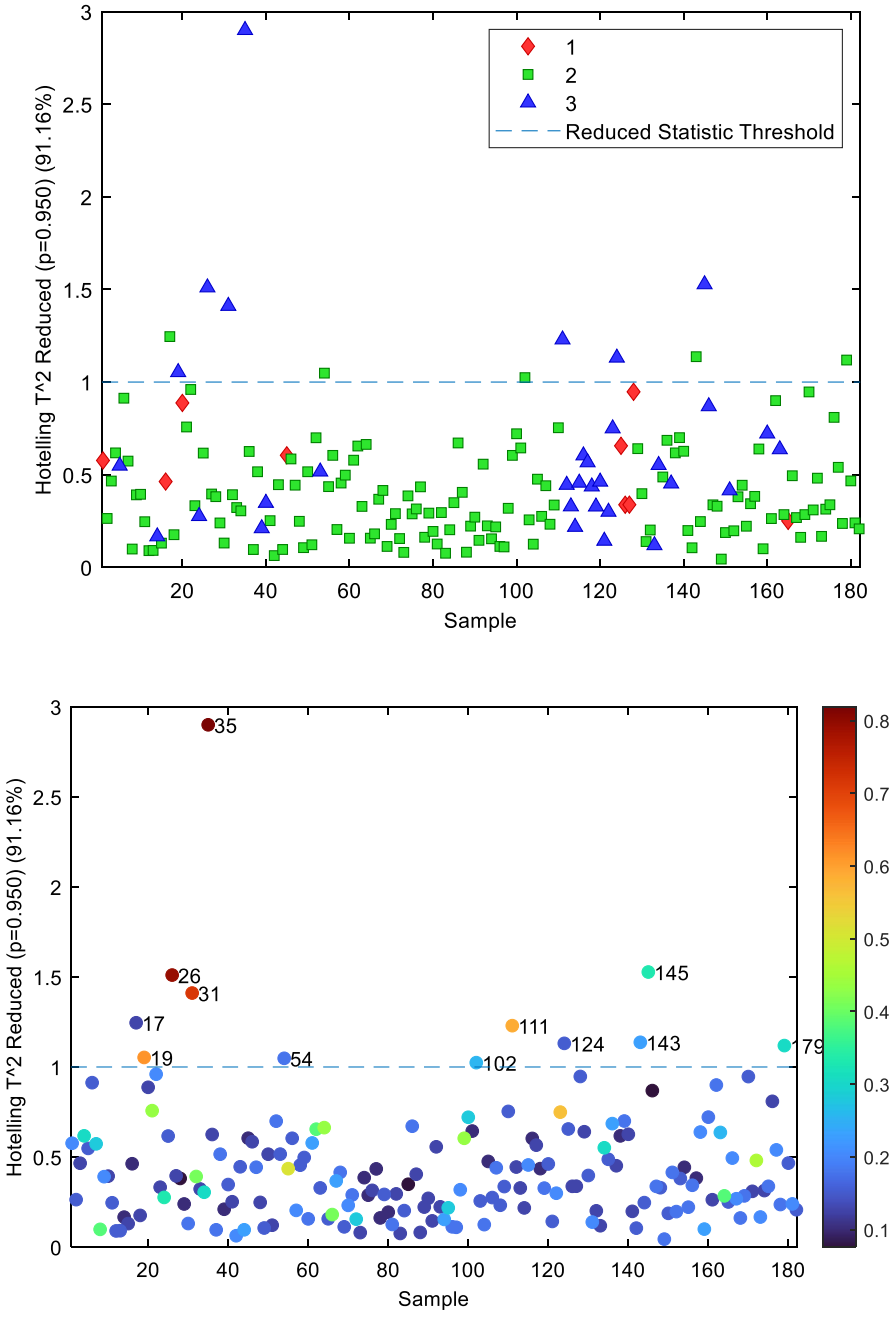


Figure 7.5. Multivariate control chart. Top samples coloured for pesto classes; bottom samples coloured for instability index.

An attempt was also made to estimate the instability index by the NIR spectra. To build the multivariate calibration model between the VIS-NIR spectra and the instability index a PLS regression model by using three latent variables (according to minimum error in cross validation with venetian blind with ten steps) was selected. The results (Figure 7.6) indicated a quite good prediction model, with a RMSECV 0.0521 and RMSEP 0.0538 indicating that NIRS could be an acceptable alternative to quicky evaluate the stability at least at pre-screening level.

This results, are especially interesting because two NIRS are already installed on-line in the Rubbiano plant in two points of the production line, and thus the predictive model could be in future implemented as real-time measure.

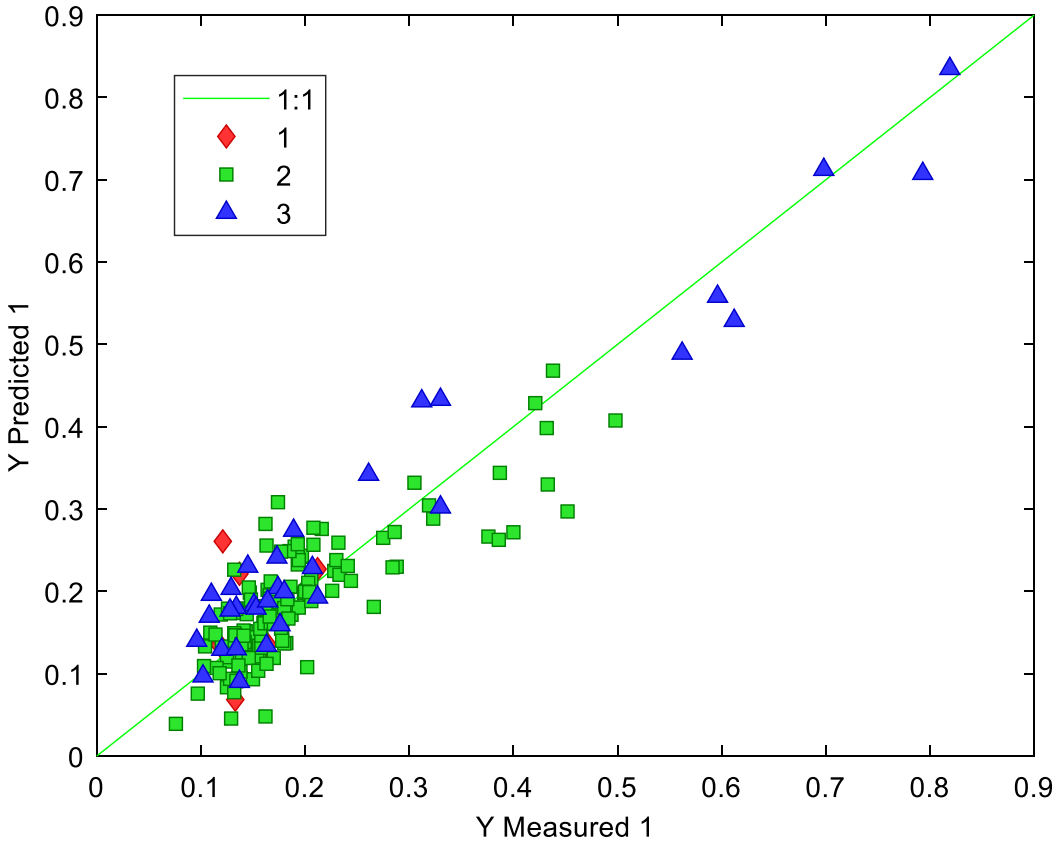


Figure 7.6. Instability index prediction by NIRS reporting the measured Instability index on x axes vs the predicted Instability index in y axes. The green line indicates the ideal prediction. Calibration and test samples are reported.

7.3.2 Results on NIR on-line on semifinished product

PCA analysis carried out on NIR spectra (acquired for 459 time points) had highlighted the presence of a cluster of samples at the negative value of PC1 and positive value of PC2, as shown in Figure 7.7a, as very far and different from all the other samples. Observing the PC1 versus time plot (Figure 7.7b), it was evident that these samples always corresponded to re-starts, where production started after a period of inactivity. In Figure 7.7c, the loadings line plots for PC1 and PC2 are shown as the blue and red lines, respectively, where it is possible to see the absorption bands as mainly responsible for this difference. However, to jointly interpret scores and loadings plots, a PC1 *vs* PC2 loadings scatter plot was also generated (Figure 7.7d). In the two figures d and c, highlighted in purple, the wavelengths that describe the separation between the NOC and anomalous samples are shown. It can be observed that the band in PC1 at 1400 nm, despite being the most intense, is not involved in the description of anomalous samples but just in extreme NOC samples with high values of PC1 scores in Figure 7.7a. On the other hand, the bands at 1170, 1213, 1236, and 1410 nm describe the behaviour of the anomalous samples, as they fell in the separation direction, meaning that these samples had quite different absorptions at these wavelengths. In detail, the bands at 1178 and 1410 nm can be ascribable to lignin, namely, the second overtone of C-H bond stretching of CH_3, and to the first overtone of the O-H bond stretching of the ROH group, respectively. Whereas the band at 1213 and 1236 nm are related to the first and second overtone of CH bond stretching of oleic and linoleic acid in olive oil CH_2 (4, 5, 6).

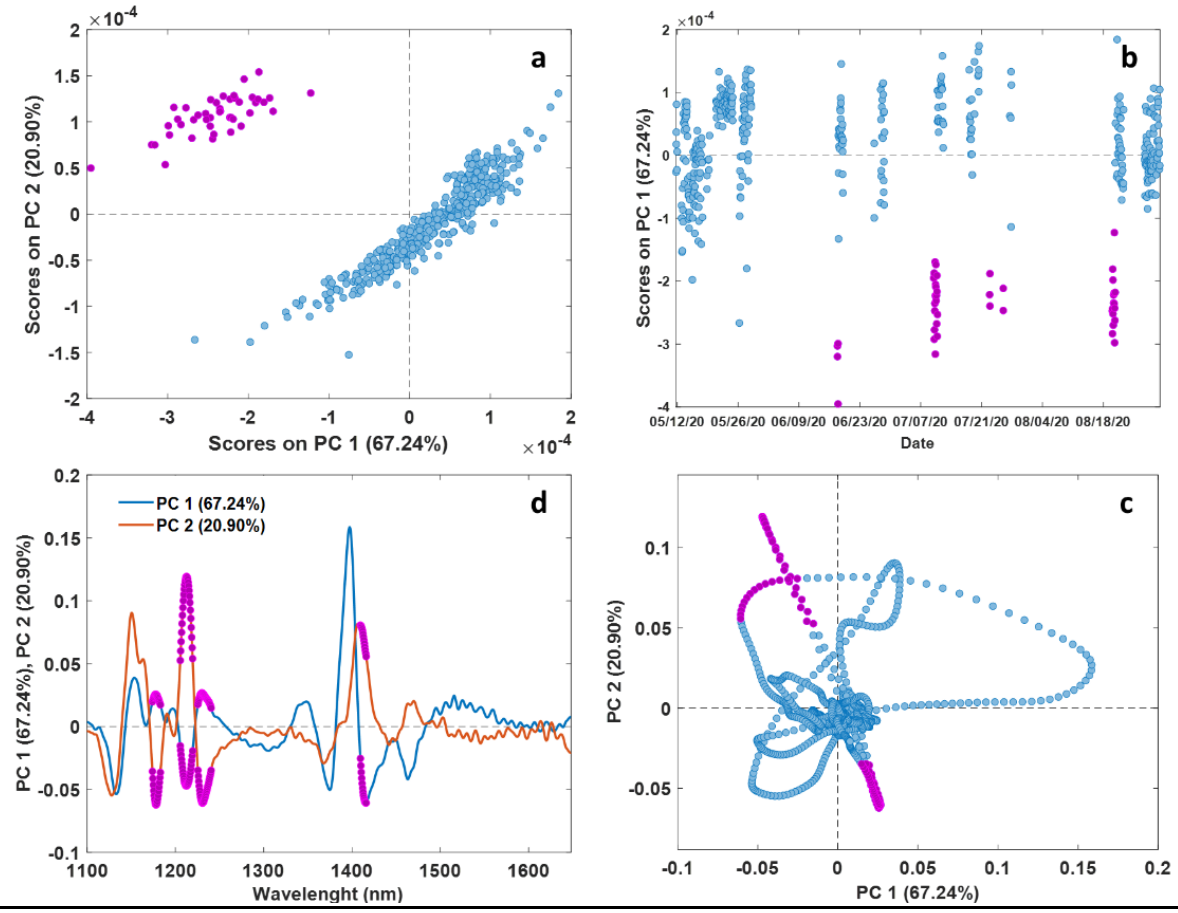


Figure 7.7. Results of the Exploratory Data Analysis performed on NIR data. PC1 vs PC2 scores plot (a), Scores on PC1 as a function of time (b), Loadings on PC1 and PC2 as a function of time (c) and Loadings on PC1 vs PC2 (d). In (a) and (b) purple points represent anomalous samples; in (c) and (d) purple points represent wavelengths that are depicting the difference of anomalous samples from the other ones.

Since these samples show outliers behaviour, as they clearly do not represent the Normal Operative Condition (NOC), were removed and a new PCA model was built to obtain a better visualization of differences among NOC samples.

The first PC (79.36% of variance explained) did not show any interesting trend, so PC2 and PC3 were inspected. In Figure 7.8a and Figure 7.8b are reported the scores plot of PC2 vs PC3 where samples are coloured according to the different additional information available i.e., suppliers and different cuts, respectively. The suppliers' names have not been disclosed because of confidential agreement restrictions. PC2 discriminates samples according to suppliers, as almost all samples of supplier number 2 have positive PC2 values and samples of supplier's number 3 and 4 have negative PC values, suggesting that they are more similar to each other with respect to number 2. Only the samples coming from supplier number 5 does not clearly differentiates from the others, whereas the number of samples of supplier number 1 is too low to judge. Furthermore, PC2 and PC3 can distinguish between samples related to cut 1 and 2 (negative values of PC2 and positive values of PC3) with respect to samples related to cut 3 and 4. The possibility to discriminate different cuts is relevant for the company, as younger basil plants generally give a higher quality product. However, observing the two plots simultaneously, it is evident that only certain suppliers, namely number 3 and 4, have delivered samples characterized by low cuts. In Figure 7.9a and Figure 7.9b are reported the loadings plots of PC2 and PC3, respectively, that show the NIR bands responsible of these differences. Even if it is not possible to assess if suppliers or cuts influence them, PCA resulted to be a valuable tool to assess if incoming information about raw material could be linked to the intermediate product characteristics, obviously a more systematic planning of the next harvesting campaigns could clarify if cut or supplier is the influential factor.

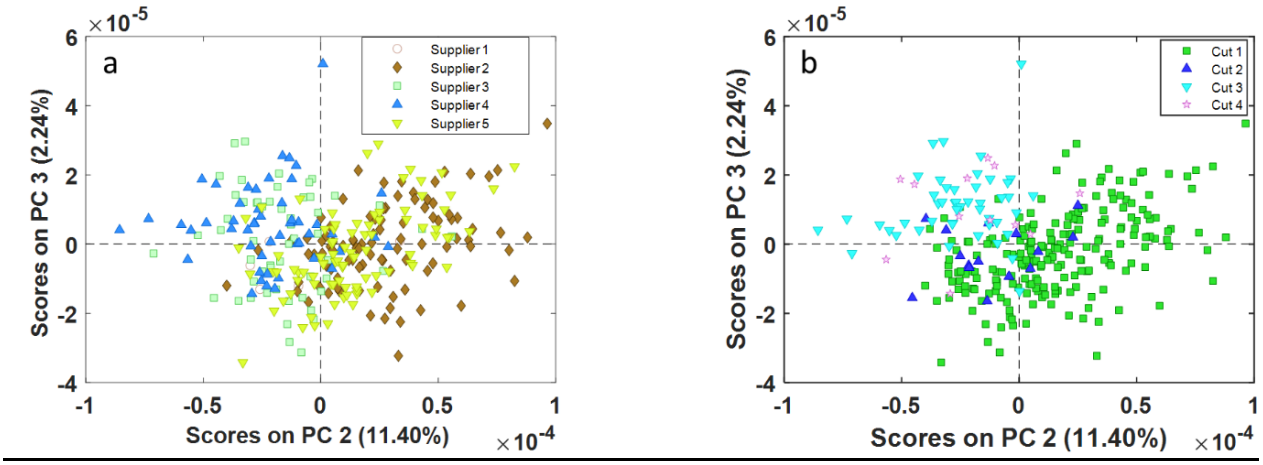


Figure 7.8. Results of the Exploratory Data Analysis performed on NIR data. PC2 vs PC3 Scores plots coloured by different suppliers (a) and cuts (b).

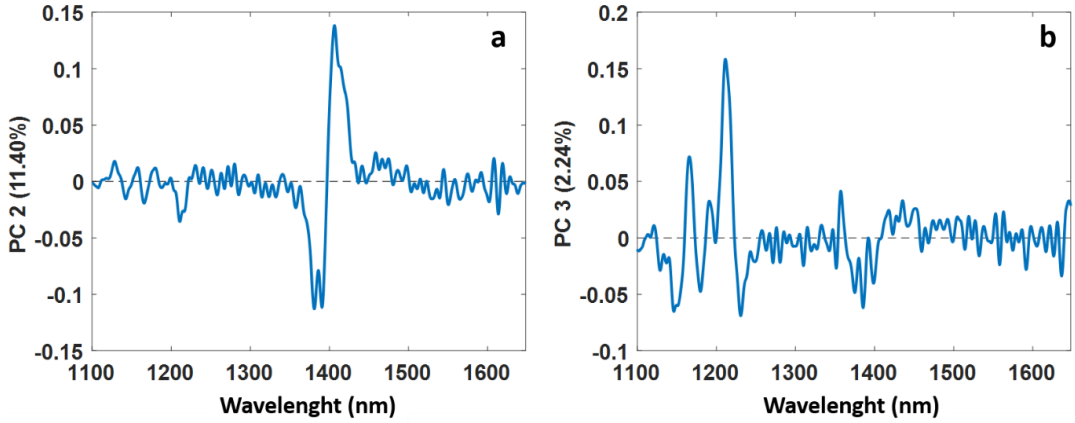


Figure 7.9. Loadings plot of PC2 and PC3, respectively.

7.3.3 On-line NIR monitoring (MSPC charts and predictive models).

As pointed out in the introduction for details refer to publication number 3.

The PCA model to build the MSPC charts, which explains 93% of the data variance with 4 Principal components, was calculated inserting in the calibration set (294 samples) only the samples that are considered in NOC according to plant experts, whereas test set (165 samples) comprised both NOC and anomalous samples. The T^2 chart, reported in Figure 7.10a, describes the distance of each sample from the origin within the model space. Black circles represent the calibration samples used to build the PCA model, whereas red diamonds represent the test samples projected on the model. This chart detected five groups of samples with high T^2 value, which again correspond to the NIR spectra acquired at the different restarts of the production. No other test sample exceeds the T^2 limit. Regarding the Q chart (Figure 7.10b), which describes the distance of each sample from the model space, the same samples corresponding to the restart are seen anomalous as for T^2 chart, meaning that the model does not describe properly these samples. Few not consecutive samples and inside the nominal 5% of the total are above the charts' limits.

Samples were also coloured according to cut, supplier, consistency, and lipids values to observe if their behaviour was related to these distinctive features, but no specific trends were detected.

Nonetheless, the results obtained show how these charts are efficient in detecting possible departure from NOC, which translate to differences in intermediate products, accelerating the identification of possible plant issues or, as in this case, the adaptation of the process while returning to NOC conditions after a stop period. NIR is a very sensible technique to signal any

variability occurring in intermediate production samples that can be due to process resetting (actual case), to process drift, or also to variation of NIR instrumentation setting/performance. Interpretation of loadings and analysis of previous production campaigns data may help discerning the different situations.

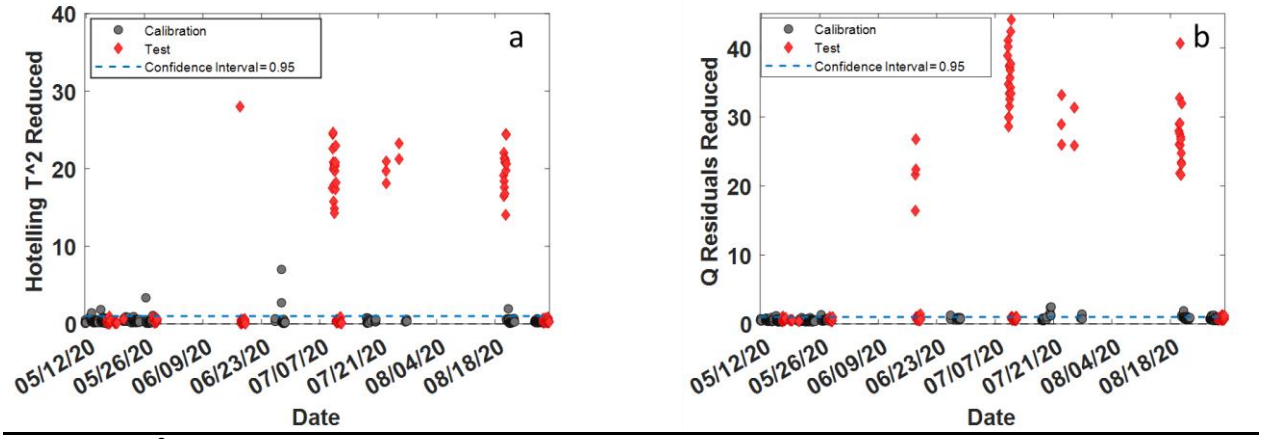


Figure 7.10. T^2- (a) and Q-(b) based MSPC charts.

NIR spectra acquired on-line can also be used to obtain multivariate calibration models to predict the quality pesto parameters, to then implement real-time prediction to get an early estimation of the pesto quality before process is finished. A main issue to take care in this case is to match the intermediate product sample (at a given production time), on which the NIR spectrum is acquired, with the correct finished pesto product at the end of production line (on which quality parameters are assessed by reference methods off-line), i.e. considering the residence time as explained in section 2.1 of paper 3.

Another, critical issue is spectral preprocessing for which the reader is referred to the section 2.4.1 of paper 3.

The main results are reported in Table 7-1.

The models for Consistency and Lipids are discussed in detail in the published paper.

The pH model shows close value of RMSECV and RMSEP, analysing the residuals plot (not shown for sake of brevity) it was possible to check the absence of bias and their random distribution, apart from 2 samples, all other samples are within a range of ± 0.15, an acceptable error for estimating the on-line quality of an intermediate product. The model commits an average percentage error of 1.1 %. The most influential bands for predicting the pH of pesto were identified through the analysis of the regression coefficients and VIP four main spectral regions were important: 1210 nm first overtone of the stretching of the C-H bond of the CH2 group of oleic and linoleic acid in olive oil and cheese; 1407 nm first overtone of the O-H bond stretching of the ROH group of lignin; 1444 nm C-H bond combination bands of the aromatic compounds in basil;1527 nm first overtone of the N-H bond stretching.

In the water activity model, the test set samples are predicted with a much higher error than the calibration set samples (RMSECV 0.001 while RMSEP 0.004). However, when analysing the residuals, a random distribution is evident and most of the samples show residue values between ±0.01, an acceptable range for the company to consider the product in specification. The model commits an average percentage error of 0.37%. Again, regression coefficients and VIPs were used to identify the most influential bands for water activity, pointing coherently to the first overtone of the O-H bond stretching (region around 1450 nm).

Finally, the dry residue model shows comparable RMSECV and RMSEP values and commits an average percentage error of 0.63%, while this is considered acceptable by the company to check whether the product is in specification for this parameter. The prediction residuals, in this

case showed a tendency to a linear trend, as if the model suffered by underfitting. The most influential bands are located at 1200 nm, 1350nm (linked to C-H bond stretching of the CH2 and CH3) and 1577 nm (first overtone of the N-H stretching of the CONH group).

Table 7-1 PLS regression results for multivariate calibration of pesto quality parameters by using on-line NIR (70/30 % calibration/validation split by duplex)

Quality parameters	***LVs***	***RMSECV***	***RMSEP***	***% average error***
Consistency (cm)	9	0.64	0.68	9.88
pH	8	0.056	0.065	1.1
Lipids (w/w%)	5	1.6	2.0	2.5
Water activity	4	0.001	0.0044	0.37
Dry residue (w/w %)	4	0.4254	0.5745	0.63

These preliminary prediction models were built as a first attempt to evaluate the possibility of predicting the properties of the final product in real time from on-line analysis of the intermediate product, showing promising results despite the limited usable data. These prediction models, being constructed with NIR spectra placed at an intermediate stage of production do not "see" the variability introduced with the addition of the final ingredients, therefore, they may not be effective when these have a strong influence on the finished product, on the other hand having a prediction albeit with a fairly significant % error but well in advance allows timely intervention in the event of estimates deviating from the required specifications. A limitation in the construction of these models is the limited range of variability of the responses which is obviously bound to the specification conditions, on the other hand there is no possibility of extending the calibration domain. Thus, these models should be viewed not so much with a view to correctly estimating the value of the property, but as tools capable of providing a preliminary assessment as to whether the tolerated ranges around the product specification are met. With this in mind, the predictive models obtained showed very good capabilities for each property, committing a percentage error acceptable to the company to consider if the product is inside specification range.

7.4 Conclusions

This part of the study presents two feasibility studies.

The first was related to the possibility to predict the pesto emulsion stability by a NIR system and demonstrate that NIR is capable to predict pesto structure instability. Its application in a production plant should be further tested.

The second part was related towards the real-time monitoring of an industrial food process line (pesto production). Since historical data were not available, the obtained results referred to a single basil harvesting campaign. The modelling effort concerned both latent variables based multivariate control charts, aimed at monitoring the stability of process conditions and the eventual detecting of fluctuations exceeding the natural variability of the process. Even though the collected data were limited, the results gave interesting insights, which are summarized below.

NIR-based multivariate control charts could detect restarts after temporary production stoppages, underlining that some changes occur in the intermediate product. On one hand, this is an indication of how sensible NIR spectroscopy is to monitor any changes, and, on the other hand, a monitoring system can clearly indication of how sensible NIR spectroscopy is to monitor any changes, and, on the other hand, a monitoring system can clearly indicate when process fluctuations return to natural process variabilities and to the constancy of the product.

The preliminary predictive models obtained showed good capabilities for each property, committing a percentage error acceptable to the company to consider the product in specification in each parameter. Their application and implementation on the line would allow in the future the early identification of intermediates that would give products that are not in specification, giving the operator the possibility of planning a verification analysis in the product laboratory (in advance of the routine scheduled time), carrying it out on specific target samples and if necessary, stopping

production or correcting it by varying the quantities of ingredients. To strengthen and validate the prediction models, it would be necessary to increase the number of samples, monitoring campaigns over few years and consider to calibrate with a mixed data set of plant and laboratory samples in order to enlarge the response variability in the calibration set. Anyhow, results show the feasibility of real-time quality monitoring to complement off-line laboratory analyses, thus reducing costs and performing quality control on all jars of pesto and not only on some samples.

---

1 Altay K, Sahingil D, and Hayaloglu AA. A geographically-registered Arapgir purple basil pesto sauce prepared with four different cheese varieties: Comparison of physical, bioactive and rheological properties. Food Chemistry Advances. 2024; 4: 100587.

2 Jackson JE, Hearne FT. Hotelling's TM2 for Principal Components—What about Absolute Values? Technometrics. 1979; 21: 253–255.

3 Nomikos P, MacGregor JF. Multivariate SPC charts for monitoring batch processes. Technometrics. 1995; 37: 41–59.

4 Galtier O, Dupuy N, Le Dréau Y, Ollivier D, Pinatel C, Kister, J et al. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. Anal. Chim. Acta. 2007; 595: 136–144.

5 Casale M, Simonetti R. Near infrared spectroscopy for analysing olive oils. J. Near Infrared Spectrosc. 2014; 22: 59–80.

6 Mailer R J. Rapid evaluation of olive oil quality by NIR reflectance spectroscopy. Journal of the American Oil Chemists' Society. 2004; 81: 823-827.

8FINAL CONCLUSIONS

8.1 Final remarks

The main objective of this PhD thesis was to improve and increase the possibility of evaluating the quality of crucial raw material and the correlated finished products. The “Pesto alla Genovese” has been a good benchmark to try several analytical and statistical approaches both in the lab and the industrial plant contexts.

Starting from the basil, (chapter 4) a crucial raw material for the pesto quality, a deep focus has been posed on the flavour characterization in an R&D context where is necessary to evaluate new basil chemotypes. The classical analytical methods used to evaluate the flavour found in the support of chemometrics a new faster, easier, and more effective way to evaluate if a flavour bouquet of a new basil chemotype is similar to a well-known basil bouquet and why it is different. It is understandable how this could help in the agronomic research.

Moreover, the study has been carried out in two modalities: with target analysis where the molecules to be measured are known, and with untargeted analysis, where no information on the expected molecules is available. The first modality is useful to control well established products, just to check their behaviour. The second modality instead is more dedicated to research situation where very new products are under exploration, as happens in R&D context.

The basil has been also evaluated for other characteristics (chapter 6) like its colour or the ratio between the leaves and the stems at the industrial production plants. This in related to the availability of a vision system that has been installed in the pesto industrial plant aiming in perspective at achieving real-time raw materials monitoring. In fact, the Quality by Desing, that is increasing in its application also in food industry contemplate that raw material should have precise characteristics. Not always easy to be sure of that with “live” row materials like basil. So, an image analysis strategy has been studied. Promising preliminary results were obtained.

In the industrial context another challenging topic has been the study undertaken to evaluate if the on-line NIR probe installed at the pesto plant monitoring a semifinished product could be, coupled to chemometrics, used to develop predictive models of the quality characteristics of the final product “Pesto alla Genovese”.

This is a very important possibility because in an industrial process having information that the final quality is not going to be the expected one, in an early stage of the process, permits to quickly intervene to correct the process.

This task has been very challenging because the production process is not as fixed as we might think, and there are pauses, stops, and restarts, minor changes in flows or in times that introduces variability difficult to control and which render the building of predictive models very challenging. Nonetheless, there are pre-processing tools which can help to study and remove the effect introduced by unwanted variability sources. On the other side, an improvement in process data storage and retrieval and automatic registration of additional information is needed.

Another relevant part of the thesis verted around the final product “Pesto alla Genovese” quality characteristics, in terms of flavour and structure.

In chapters 5 and 7 some methods have been tested with chemometrics support to evaluate both the aroma and the structure stability. Results indicate promising possibilities for techniques like NIR that has more potentiality respect to the other techniques used to be exported in a routine quality control context for its easiness of use.

Innovative analytical approaches like hyperspectral imaging (chapter 6) have been tested on Pesto to evaluate new possible tools to be exported from a research environment to a quality control lab. Also in this case, the chemometrics is essential to manage the complexity of this data.

8.2 Future perspectives

The future of the quality design is in the direction on what the Industry 4.0 expect, with the number of sensors increasing along all the production processes, connected to huge databases where big data will be then properly managed and elaborated to properly drive the production process.

On the other hand, in the Laboratories, the develop of more sophisticated techniques (i.e. GCxGC-MS or the multi-sensors hyphenated techniques), will require more and more sophisticated way to analyse the data.

This is a challenging and uphill path that will be done step by step. The explored possibilities of evaluation quality parameters done in this thesis are steps in this complex path.

8.3 To conclude

The initial idea to develop a “chemometric toolbox” to be applied in my everyday job as chemist and researcher has been successfully reached. All the studied cases have been approached with chemometric mindset and tools, that increased, or in some cases made possible, the extraction of information, the elaboration of complex data and at the end a clearer understanding of the studied topic.

It has been interesting to me to observe how wide are the possible applications of the chemometric tools: from the classical laboratory data obtained by largely used gas-chromatographs, that increase their descriptive capabilities when enhanced by chemometric, to techniques like the hyperspectral imaging, more and more used for its powerful capacity to join the morphology description of the image to punctual chemical information.

And leaving the R&D laboratory to move to the production plant, the power of chemometrics became more relevant considering the huge quantity of collected data in a context rapidly moving towards the Industry 4.0, with the spread of sensors and measure instruments.

I know that for me this journey into the chemometrics world is just started, but I have now more awareness of the huge possibilities of its use and of the pitfalls that are always round the corner.

ACKNOWLEDGEMENTS

First, I will say a great thank you to Prof. Marina Cocchi for her passion, her patience and all the time she dedicated to me.

Thanks to Prof. Caterina Durante for her suggestion and help in my doubts.

A warm thank to all the team of Prof. Cocchi: Lorenzo, Alessandra, Samuele and Daniele.

Daniele thank you for the period abroad we spent in France. It has been a beautiful experience.

Another great thank goes to Prof. Jean Michel Roger and all his team for hosting me in ChemHouse in Montpellier (France). His deep knowledge and his vitality have been an inspiration for me. Thanks to Silvia Mas-Garcia for the fruitful conversations we had in ChemHouse.

I will also thank my Company, in the persons of dott. Rosamaria Petrosino and dott. Francesca Vitali, that trusted me and believed in this project.

And to finish, I would like to thank my wife Lidia for all the time and weekends spent away from her studying.

PUBLISHED PAPERS

In Appendix 1 all the published articles.

PAPER 1

molecules

MDPI

Article

Characterization of Basil Volatile Fraction and Study of Its Agronomic Variation by ASCA

Alessandro D'Alessandro [1], Daniele Ballestrieri [1], Lorenzo Strani [2], Marina Cocchi [2,*] and Caterina Durante [2]

[1] Barilla G. e R. Fratelli, Via Mantova 166, 43122 Parma, Italy; alessandro.dalessandro@barilla.com (A.D.); daniele.ballestrieri@barilla.com (D.B.)
[2] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy; lostrani@unimore.it (L.S.); cdurante@unimore.it (C.D.)
* Correspondence: marina.cocchi@unimore.it; Tel.: +39-059-2058-554

Abstract: Basil is a plant known worldwide for its culinary and health attributes. It counts more than a hundred and fifty species and many more chemo-types due to its easy cross-breeds. Each species and each chemo-type have a typical aroma pattern and selecting the proper one is crucial for the food industry. Twelve basil varieties have been studied over three years (2018–2020), as have four different cuts. To characterize the aroma profile, nine typical basil flavour molecules have been selected using a gas chromatography–mass spectrometry coupled with an olfactometer (GC–MS/O). The concentrations of the nine selected molecules were measured by an ultra-fast CG e-nose and Principal Component Analysis (PCA) was applied to detect possible differences among the samples. The PCA results highlighted differences between harvesting years, mainly for 2018, whereas no observable clusters were found concerning varieties and cuts, probably due to the combined effects of the investigated factors. For this reason, the ANOVA Simultaneous Component Analysis (ASCA) methodology was applied on a balanced a posteriori designed dataset. All the considered factors and interactions were statistically significant ($p < 0.05$) in explaining differences between the basil aroma profiles, with more relevant effects of variety and year.

Keywords: basil; aroma; fast GC; GC/O; electronic nose; PCA; ASCA; cut; variety

Citation: D'Alessandro, A.; Ballestrieri, D.; Strani, L.; Cocchi, M.; Durante, C. Characterization of Basil Volatile Fraction and Study of Its Agronomic Variation by ASCA. *Molecules* **2021**, *26*, 3842. https://doi.org/10.3390/molecules26133842

Academic Editors: Alessandra Biancolillo and Angelo Antonio D'Archivio

Received: 28 April 2021
Accepted: 18 June 2021
Published: 24 June 2021

1. Introduction

Basil (*Ocimum basilicum* L.) is an annual plant of the *Lamiaceae* family, known worldwide as a culinary and healthy herb [1]. Basil's essential oils have been used in many fields for medicinal treatments, perfumery and cooking spices. Originating from India, Africa and Asia, its cultivation is now spread worldwide [2].

It is estimated that basil counts from fifty to one hundred fifty species, of which the most commonly used in the culinary field is sweet basil [3,4]. It is present in many different chemo-types due to its characteristic to easily cross-breeds [4,5]. For that reason, it can sometimes be challenging to determine the species or the variety of a basil plant. Its characteristics in terms of morphology, agronomy performances and aroma pattern are normally determined [6–8]. These characteristics are influenced not only by the chemotype/species/variety, but also by agronomic practices, climatic conditions and age of the plant [1,3].

The basil aroma is composed of a large number of molecules, mainly terpenoids, alcohols, aldehydes, ketones and esters [3,9]. Totally, there are more than one hundred molecules, of which the most representatives in sweet basil are considered linalool, estragole, eugenol and eucalyptol (1,8-cineole) [7,10]. The content of these molecules could give a preliminary evaluation of different basil flavour profiles, while a more accurate evaluation of the final aroma will also consider the concentrations of other minor components, mainly the molecules that have a low odour threshold [11,12]. The odour threshold is defined as the lowest concentration of a molecule that could be perceived by olfaction.

Thus, in the evaluation of the flavour patterns, it is necessary to consider not only the concentration of a given molecule but also its capacity to be perceived.

Basil is one of the main components of the "Pesto Genovese" sauce, a typical and well appreciated Italian green sauce. The basil aroma pattern is crucial for the organoleptic features of pesto sauce and consequently its analytical characterization is relevant [13] in terms of selecting the preferred profile or to search for new patterns.

There are many different methods to identify and quantify volatile organic compounds (VOCs) based on gas chromatography (GC) and mass spectrometry (MS), either coupled or not, and using different systems of sampling. Among coupled GC–MS methods, different systems are available to collect, trap and concentrate the VOCs, such as headspace solid phase microextraction gas chromatography–mass spectrometry (HS-SPME-GC–MS) [6], headspace sorptive extraction gas chromatography–mass spectrometry (HSSE GC–MS) [13], dynamic headspace-thermal desorption–gas chromatography/mass spectrometry (DH-TDU-GC–MS) [14]. Direct-injection mass spectrometry (DIMS) [15], without a separation step, is also very diffuse in food analysis [16–18]. In particular, the development of an ambient ionization mass spectrometer (AMS) [19–23] is very important, especially coupled with the development of miniature and portable mass spectrometers [21–23] and innovative introduction systems, such as membrane inlet mass spectrometers (MIMS) [21,24]. AMS, while opening up very interesting perspectives for in situ food analysis and control, has still to become an established reference for quantitative analysis, especially for solid samples [19].

The basil aroma pattern, to the best of our knowledge, has been characterized only by GC based techniques, for instance, headspace solid phase microextraction gas chromatography–mass spectrometry (HS-SPME-GC–MS) [6], headspace sorptive extraction gas chromatography–mass spectrometry (HSSE GC–MS) [13], as well as gas chromatography as such (GC and GC–MS) [10], indirectly measuring the total phenolic compounds [25] or using flow-injection mass spectrometry [18].

As basil is a very delicate plant, which is difficult to store after cutting [26,27], it would be extremely useful to have a fast analytical method, being at the same time suitable to discriminate the different varieties and furnishing information on the compositional profile of the aroma fraction.

To this aim, in this paper, we tested an electronic nose system based on ultrafast gas chromatography (fast-GC) since it can provide a non-invasive, rapid, sensitive and relatively low-cost system. Moreover, it allows direct comparison with sensory evaluation that is usually carried out by gas chromatography–olfactometry (GC/O) [28] analysis. In particular, the Heracles II e-nose device [29] was tested, which has been previously applied to characterize the volatile fraction of different food commodities [30–33], while there is, to the best of authors' knowledge, no study concerning basil or other spices. The aroma profile gathered by fast-GC was matched with sensory evaluation from GC/O, and the detected molecules, mainly perceived in the basil flavour pattern and persistent in GC/O, were quantified.

The developed methodology was applied to evaluate several basil varieties, grown on open fields in different years considering more cuts, to obtain a preliminary overview by multivariate exploratory data analysis of the aroma variation due to both varieties and period of harvesting. A deepest insight and a better understanding of these effects can be gathered by ANOVA–Simultaneous Component Analysis (ASCA) [34], which generalizes classical analysis of variance (ANOVA) to multivariate data, overcoming the main limitations (number of samples higher than number of variables, breakdown in case of variables collinearity) and multinormal distribution assumption of multivariate ANOVA (MANOVA). First, a classic ANOVA was carried out to split the data matrix into the effect matrices for each experimental factor and their interactions. Then, simultaneous component analysis was carried out on the effect matrices to identify and visualize the contribution of the measured variables to each of the effects that introduced systematic variation [35]. One of the main advantages of ASCA is the interpretation of the factor

levels in terms of the measured variables through loadings inspection. ASCA has been successfully applied in metabolomics [34,35], as well as in food analysis [36–38].

ASCA requires data coming from an experimental design, and thus we applied it to a balanced reduced set of varieties, in order to investigate the effects of cutting period, basil variety and harvesting year on the basil aroma pattern.

2. Materials and Methods

2.1. Basil Plants

The plants of basil (*Ocimum basilicum*) of twelve different commercial varieties of "genovese" type were supplied, for all the samples, by local producers (Parma Vivai). The varieties name is indicated with a code for confidentiality reasons. Only the "Italiano Classico" has been indicated because it is largely commercially used. All plants have been grown in open fields following standard agricultural practices. Each basil variety was collected at different plant ages: in most cases two cuts were collected and sometimes up to four cuts were taken (Table 1). Plants were cut leaving about 5–6 cm from soils, allowing the plant to regrow for the next cut. The first cut was carried out when the plants were aged 40 days, while the subsequent cuts were carried out at time intervals of about 20 days each. Finally, in order to have a preliminary idea on the variation of the investigated aroma fraction as a function of the harvest, different basil varieties were collected for three years (2018–2020). In Table 1, the number of samples per year, variety and cut are reported.

Table 1. Samples analysed in the three years of experiment with the indication of the samples undertaken for each cut.

Crop Year	Basil Variety	Cut in Bold (No. of Samples)
2018	italiano classico	**1st** (11), **2nd** (12), **3rd** (3)
	variety 3	**1st** (1), **2nd** (1)
	variety 5	**1st** (1), **2nd** (1)
	variety 8	**1st** (1), **2nd** (1)
	variety 10	**1st** (1), **2nd** (1)
	variety 11	**1st** (1), **2nd** (1)
2019	italiano classico	**1st** (4), **2nd** (2), **3rd** (2), **4th** (2)
	variety 5	**1st** (2), **2nd** (1), **3rd** (1), **4th** (1)
	variety 8	**1st** (2)
	variety 9	**2nd** (1), **3rd** (1), **4th** (1)
	variety 10	**1st** (2), **2nd** (1), **3rd** (1), **4th** (1)
	variety 11	**1st** (2), **2nd** (1), **3rd** (1), **4th** (1)
2020	italiano classico	**2nd** (2), **3rd** (1), **4th** (2)
	variety 1	**2nd** (1), **3rd** (1), **4th** (1)
	variety 2	**2nd** (1), **3rd** (1), **4th** (1)
	variety 4	**2nd** (1), **3rd** (1), **4th** (1)
	variety 5	**2nd** (1), **4th** (1)
	variety 6	**3rd** (1), **4th** (1)
	variety 7	**2nd** (1), **3rd** (1), **4th** (1)
	variety 9	**2nd** (1), **4th** (1)

2.2. Sample Preparation

Basil plants were collected early in the morning, typically from 4 to 8 a.m., and were immediately sent to the lab for the evaluations. Plants were analysed within 6–8 h from the cut. About 30 g was exactly weighted at 0.1 g of the whole basil plant, including leaves and stems, and was hashed in a blender (Oster, Sunbeam Products Inc., Boca Raton, FL, USA) for 30 s in 300 mL of extraction solution at room temperature. The extraction solution was prepared with NaCl at a concentration of 100 g L^{-1}, to increase the volatiles release in the headspace (next step of the analysis), and 6 mg kg^{-1} of ethyl iso-butyrate to serve as internal standard for the CG analysis. After 30 s of resting time, 20 µL of the solution was collected and transferred in 20 mL amber vials that were immediately sealed and sent

for analysis. Each extract was sampled at least three times in different vials. All reagents, standard and solvents were analytical grade (Sigma Aldrich, Inc., Saint Louis, MO, USA).

2.3. Heracles e-Nose Analysis

The analysis of the volatile molecules in the sample headspace was carried out using a Heracles II (Alpha MOS, Tuluse, France) ultra-fast chromatography electronic nose [18]. The instrument consists of a double-columns ultra-fast-chromatography system, with FID detectors, interfaced with a PAL-RSI automatic headspace autosampler, after injection a Tenax TA polymer trap is employed. The two columns were mounted in parallel in the oven; they had different polarities, namely, an MXT-5 (non-polar) and MXT-1701 (slightly polar) were employed, both 10 m in length, with internal diameters of 0.18 mm and phase thicknesses of 0.40 µm. A temperature ramp was employed, starting from 50 °C for 2 s, then going to 80 °C at 1 °C·s^{-1} and finally reaching 250 °C at 3 °C·s^{-1}. The total fast GC analysis time was 110 s. The carrier gas was hydrogen.

The different replicates of each extracted sample were loaded in the instrument autosampler and incubated for 20 min at 40 °C before injection with 500 rpm agitation (5 s on, 2 s off). Then, 1 mL of air headspace was injected with a syringe temperature of 50 °C. Trap loading conditions were 18 s at 40 °C, then flashed to 250 °C for the release in the two columns at split ratio 1:1.

The AlphaSoft v 16.0 software was used to process the data. Volatile compounds were identified on the basis of Kovats' relative retention indices (KI) and can be linked to specific molecules that are collected in the AroChemBase v 7.0 database (Alpha MOS., Tuluse, France). In this way, eighteen compounds were tentatively identified as further discussed in Section 3.

2.4. Gas Chromatography–Mass Spectrometry Olfactometry Analysis (GC–MS/O)

To select the key molecules perceived in basil aroma, a preliminary analysis on the Italiano Classic variety was conducted by gas chromatography–mass spectrometry coupled with a Gerstel ODP3 sniffing port olfactometer (GC–MS/O). Among the about one hundred and fifty molecules observed in GC–MS (data not shown), only thirty-two were perceived by GC/O sniffing trained panellists in terms of odour, and just nine of these had shown a persistent odour after three dilution steps. Matching these molecules with the eighteen molecules observed in the Heracles chromatograms, nine key marker molecules were selected as the most representative of the basil flavour pattern, as reported in Table 2.

Table 2. Persistent molecules found in basil aroma, selected by GC/O, with CAS Number and the descriptions assigned by the CC-O panelists.

Molecules	CAS Number	Aroma Description
hexanal	66-25-1	green grass, rancid
2-hexenal	63449-41-2	spices/herbal
a-pinene	80-56-8	herbal, woody
b-myrcene	123-35-3	flower, cytrus
eucalyptol	470-82-6	balsamic, eucalyptus, menthol
linalool	78-70-6	flower, cytrus, vinegar
estragole	140-67-0	anis, liquorice, fennel
eugenol	97-53-0	cloves, spices
b-caryophyllene	87-44-5	spices

2.5. Quantification of Key Molecules

For each of the key nine molecules of interest, a calibration curve was obtained (Table 3) by preparing standard solutions at six concentration levels, using ethyl iso-butyrate as internal standard. Two mother solutions were prepared. First, a solution of ethyl iso-butyrate (internal standard) was prepared by diluting about 100 mg in ethanol, exactly weighed, in a 100 mL volumetric flask to obtain a final concentration of about 1000 mg kg^{-1}.

The second solution of multistandards was prepared by diluting, in 100 mL of ethanol, quantities of each standard exactly weighed from 60 to 150 mg, depending on the respective standard volatility, to obtain final concentrations ranging from 600 to 1200 mg kg^{-1}. The six calibration solutions at different level concentrations were prepared by diluting with ethanol to 5 mL final volume, 0.5 mL of the IS solution and, respectively 0.25, 0.5, 1.0, 1.5, 2.0 and 3.0 mL of the multistandards solution. From each calibration solution, 1 µL was collected and loaded into the 20 mL vials for the analysis. The calibration curve was obtained normalizing the area of each analyte with respect to the internal standard area and quantity. A representative chromatogram for one of the multistandards solutions used for calibration is shown in Figure 1.

Table 3. Coefficient of determination (R^2), slope of the calibration curves, and limit of detection for the investigated compounds.

Compounds	R^2	Slope ± SD	LOD (µg kg^{-1})
hexanal	0.9997	0.96 ± 0.01	47
2-hexenal	0.9998	0.79 ± 0.01	23
a-pinene	0.9998	1.73 ± 0.01	28
b-myrcene	0.9999	1.61 ± 0.01	11
eucalyptol	0.9999	1.88 ± 0.01	22
linalool	0.9995	0.394 ± 0.004	60
estragole	0.9994	1.33 ± 0.02	52
eugenol	0.9999	0.453 ± 0.002	32
b-caryophyllene	0.9968	1.22 ± 0.03	22

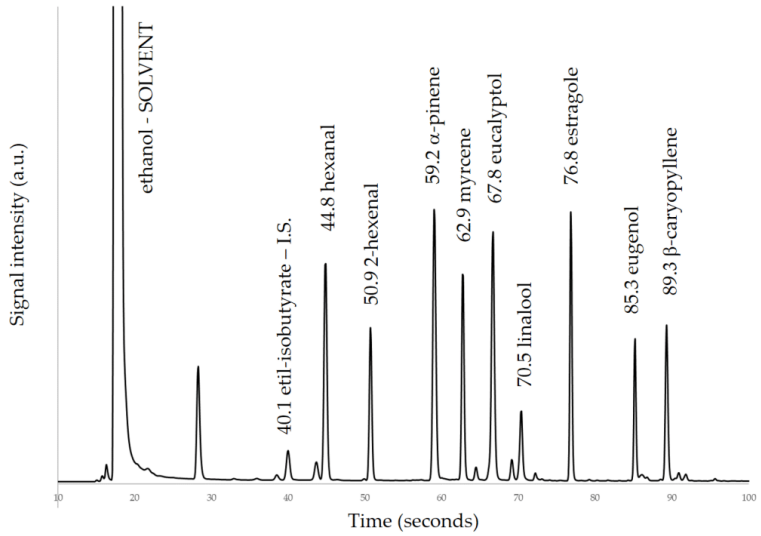


Figure 1. Chromatogram of multistandards solution. The peaks of the nine molecules with their retention times are shown, together with the peak of internal standard (IS) and solvent. Peak just before 30 s and other minor peaks are solvent impurities.

As far as the nine investigated compounds are concerned, the calibration curves were linear over the examined concentration range. In Table 3, the coefficient of determination, the slope and the limit of detection (LOD) for each calibration curve are reported.

At the start of a new analytical batch, three empty vials were injected as blanks to clean the system and one empty vial was run between each group of replicates of the samples to assure the system was clean and prevent cross-contaminations between different samples.

The concentration of each molecule was calculated with respect to the exact weight of the plant basil extracted. As a result, a dataset of the concentration in μg kg^{-1} of all the nine marker molecules of the basil samples was obtained.

In order to evaluate the short-term (intra-day) and long-term (inter-day) reproducibility, nine replicates of the same basil sample were prepared from scratch and analysed in the same day at different times, and in three different days, respectively. The relative standard deviation (RSD) was then computed for both reproducibility conditions. In particular, intra-day RSD ranged between 4 and 9%, while inter-day RSD ranged between 8 and 10%, showing good reproducibility values.

2.6. Data Analysis

Principal Component Analysis (PCA) was performed on the obtained concentration dataset (267 × 9), composed of the samples reported in Table 1, including the three replicate extracts (Section 2.2) for each sample. The samples varied according to three factors: year of cultivation (2018–2020), cut (1st, 2nd, 3rd and 4th) and basil variety (12 varieties).

Data were autoscaled to allow each of the nine molecules to contribute to the model independently of being a major or minor component.

ANOVA–Simultaneous Component Analysis (ASCA) method [24] was used to evaluate the potential significance of the effect of the three above-mentioned factors and their interactions. ASCA performs a classical ANOVA, partitioning the variability of the data into the contribution of each factor and interaction:

$$X_c = X - 1m^T = X_1 + X_2 + X_3 + X_{1x2} + X_{1x3} + X_{2x3} + X_{1x2x3} + X_{res} \quad (1)$$

where X is the scaled data matrix, m^T is the mean profile of the samples, X (1, 2 and 3) are the matrices related to the main effects, and X (1 × 2, 1 × 3, 2 × 3 and 1 × 2 × 3) are the matrices linked to the interaction effects. The rows of these matrices are highly structured, e.g., all rows related to one level (as an example 2019, for the factor year) are equal in X_1 and analogously all rows of X_2 and X_3 are equal for each cut and type of variety. Interaction matrices also have equal rows for the same level of interaction. X_{res} hold the residuals.

Then, each matrix was analysed by a distinct PCA model and Equation (1) can be reformulated as:

$$X_c = T_1P_1 + T_2P_2 + T_3P_3 + T_{1x2}P_{1x2} + \cdots + X_{res} \quad (2)$$

where T holds the scores and P the loadings of each PCA model, the maximum number of PCs for each model is equal to the number of levels minus one.

In order to better inspect the ASCA results, i.e., to highlight how the samples are dispersed around the mean of each effect level, it is useful to project the single samples on the ASCA scores plot. This can be achieved by adding the residuals to the estimated x_i values and then calculating the single sample scores, i.e., for each factor or interaction (f), a computation of the score vector $t_{i+res}(f)$ has been carried out through the following equation:

$$t_{i+res}(f) = (X_i(f) + X_{res})p_{res}(f) \quad (3)$$

where $X_i(f)$ is the effect matrix for a specific factor or interaction and and X_{res} is the residuals matrix, whereas $p_{res}(f)$ represents the loadings vector of the SCA model for the effect of that factor or interaction.

Since ASCA requires a balanced design of experiments to work properly, just 12 different conditions were selected from the whole dataset, leading to a total of 36 experiments as shown in Table 4. In fact, at the beginning of experimentation, a balanced design was not undertaken, also due to the limited availability of varieties which could be cultivated by the single producers; thus, it was not possible to study all levels for each of the experimental factors.

Table 4. Design of experiments structure for ASCA.

Year	Cut	Variety
2019	2	Variety 5
2019	2	Italiano Classico
2019	2	Variety 9
2019	4	Variety 5
2019	4	Italiano Classico
2019	4	Variety 9
2020	2	Variety 5
2020	2	Italiano Classico
2020	2	Variety 9
2020	4	Variety 5
2020	4	Italiano Classico
2020	4	Variety 9

Therefore, a balanced a posteriori design was built considering two levels for the factors "year of cultivation" (2019 and 2020) and "cut" (second and fourth) and three levels for the factor "variety" (Italiano Classico, Variety 5 and Variety 9). The significance of the effect of each design factor or interaction was assessed by means of permutation tests with 1000 randomizations [39,40].

Software

Data analysis was performed using routines and toolboxes developed in the Matlab 2020b environment (the Mathworks Inc., Natick, MA, USA). Principal component analysis has been carried out by PLS-Toolbox v. 8.9 (Eigenvector Inc., Manson, WA, USA). ASCA has been carried out by using routines developed and kindly made available by Dr. F. Marini, University of Roma La Sapienza (Italy).

3. Results and Discussion

3.1. Aroma Analysis

The pattern of volatile compounds of basil highlighted by the fast-CG analysis comprises eighteen molecules that were tentatively identified by using the Kovats relative retention indexes. The Heracles software compares the retention indexes of the two columns of different polarities to improve the tentative identification. In Figure 2, the identified molecules are shown. Among them, there are the nine ones that were identified as relevant in terms of persistent perceived odour, thus indicating that the fast-CG technique is suitable to characterise basil aroma.

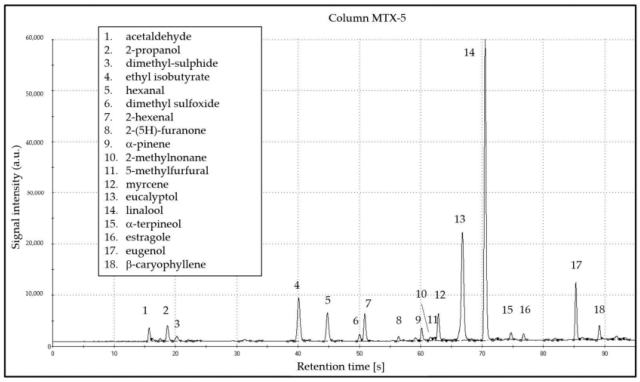


Figure 2. An example chromatogram obtained by elution on column MXT-5 of Heracles II. Peak 4 is the internal standard.

The identification of these nine molecules was confirmed by comparison with the elution time of injected standards and, once quantified, their concentrations were consistent with a typical "eucalypt" basil volatile pattern [6,8] with the prevalence of linalool, followed by eucalyptol (1,8-cineole) and then by eugenol. Other molecules are typical of essential oils of basil such as hexanal, α-pinene, myrcene and caryophyllene [41].

As previously reported [7], the flavour profile is strictly related to the presence or the prevalence of key odorant molecules, with a consequent impact on the final perceived bouquet. Four main basil chemotypes have been described by Lawernce et al. [42] depending on the prevalence of odorant molecules: estragole rich, linalool rich, methyl-eugenol rich and methyl cinnamate rich. Varieties used in the present study held predominantly in the linalool rich chemotype, but with some diversity. Variety 8, for example, was characterized for its lower level of linalool compared to other varieties, whereas on the contrary, variety 9 had the higher value. In a similar way, estragole was relatively more present in varieties 8 and 9 with respect to other varieties.

3.2. Multivariate Exploratory Analysis

PCA analysis was applied to the autoscaled data matrix composed by the nine volatile molecules obtained for the 267 samples characterized by different varieties, cuts and harvested years. Autoscaling was selected as the most appropriate data preprocessing method as the different volatile compounds had different variances due to their different concentration ranges. In this first exploratory analysis, two principal components seemed appropriate considering their explained variance (Figure 3).

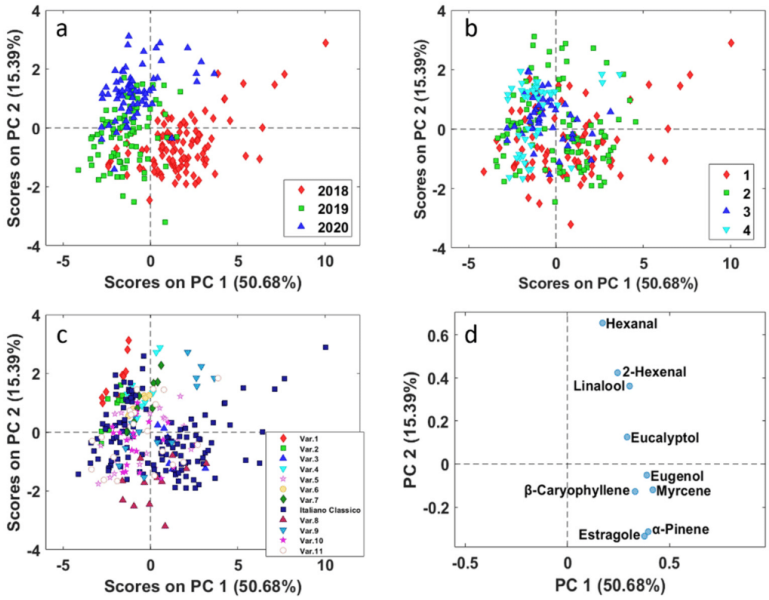


Figure 3. PCA of all basil samples (Table 1). PC1 vs. PC2 scores (**a**–**c**) and loadings (**d**) plots. Basil samples are coloured according to: (**a**) year; (**b**) cut; (**c**) variety.

In Figure 3, the PC1 vs. PC2 scores plot is reported and the different basil samples are represented with different symbols and colour according to year (Figure 3a), cut (Figure 3b) and basil variety (Figure 3c).

From the PCA results some information could be obtained. In particular, Figure 3a shows that slight differences could be observed among the three harvesting years, more in

2018 than in 2019 and 2020. The main contribution to this separation seems to be given by a higher concentration of almost all the investigated volatile molecules, since they lie on the same side of the respective loadings plot, all at positive values (Figure 3d). This difference is within the expected yearly variability, due to the different weather conditions. As an example, the year 2018 was probably characterized by less rainfall than in the years 2019 and 2020. As far as different basil cuts are concerned, Figure 3b points out that well defined clusters are not observable with respect to different basil cuts. Cut number 4, located on the left of the scores plot, is more homogeneous, at first it seems that the average level of all the flavour molecules is lower than in the other cuts; however, this information overlaps with that of the year.

In Figure 3c, the different varieties are rather overlapped, and it is evident a "spread" of Italiano Classico basil variety samples, which are uniformly distributed along the variability range of the scores space. Notwithstanding, PC2 highlights the difference of basil variety 8, which has the most negative scores on PC2 and thus presents a higher value of estragole and alfa-pinene (negative loadings values on PC2). A few samples harvested in 2020 of varieties 1, 4 and 9, and of Italiano Classico harvested in 2018, show high positive scores value on PC2, corresponding to higher amount of hexanal (most positive loadings value on PC2), whose odour is described as "green grass", and could give, depending on its concentration, an unwanted "hay" note.

Finally, it can be observed that varieties 1, 2, 4, 6 and 7, which were cultivated only in 2020, are mostly located in the first quadrant (negative PC1 and positive PC2 score values) this indicates a lower amount of estragole, alfa-pinene, myrcene, b-caryophyllene, and eugenol, which fall in the opposite quadrant in the loadings space (positive PC1 and negative PC2 loading values) and thus less fruity/floral and spicy odours.

In general, the interpretation of the overall PCA results is hampered due to the combined effects of all the investigated factors.

For these reasons, after this preliminary investigation, ASCA methodology was applied on the balanced reduced dataset (Table 4) with the aim to assess if the considered experimental factors and their interactions could have a significant effect on basil's aromatic profile. As a first step, ASCA performs a partition of the data variability into the contribution of each factor and interaction. In this case, the variation of the original data matrix was partitioned in eight different submatrices: three describing the main effect of each experimental factor—year, cuts and variety; three corresponding to the effect of each second order interaction (any possible combination of levels for each couple of factors); one accounting for the three-way interaction effect (not considered in this study), and one holding the residuals. The significance of the factors or interactions' effects was assessed by means of a permutation test, which compares the experimental sum of squares for each effect matrix with its corresponding distribution under the null hypothesis. Results of the test are shown in Table 5, where the explained variance and probability p-value are reported for each factor and their second order interaction. All the considered factors and interactions were statistically significant ($p < 0.05$), even though the effect of the factors "variety" and "year" presented a higher explained variance than other effects. On the other hand, the effect of factor "cut" explained just 3% of the total variance, suggesting a lower influence on basil's aromatic profile compared with the other two main factors. This can also be seen in the fact that the second order interactions in which factor "cut" is involved explain less than the 4% of the total variance, whereas the interaction "year $\times$ variety" explains almost 12%.

Table 5. Explained variance and probability values for main factors and their second order interactions.

Factor	Expl. Var. %	*p*
Variety	36.41	<0.001
Year	22.31	<0.001
Year × Variety	11.95	<0.001
Year × Cut	3.74	<0.001
Cut × Variety	3.1	0.003
Cut	3	<0.001

After the assessment of the significance of each factor and interaction, a component analysis (SCA) was performed on each effect matrix separately in order to interpret the observed variation. In Figure 4a, the scores plot of the effect for factor "year", with projected residuals, is shown. This plot was obtained according to Equation (3). Since the year effect matrix contains just two rows, one for each considered year, the SCA model is described by only one component (SC1), which explains 100% of the variance.

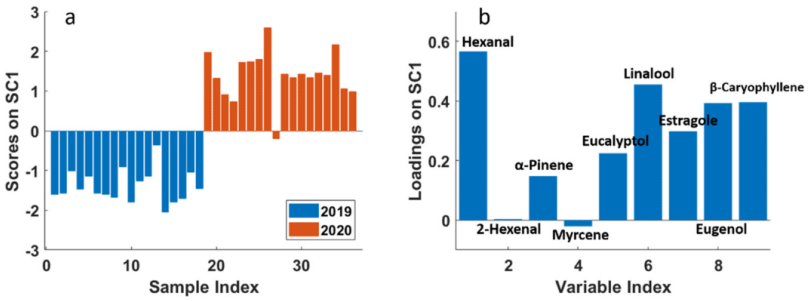


Figure 4. SCA of the effect matrix "year". (**a**) Scores plot (SC1) with projected residuals; (**b**) variable loadings (SC1).

From the scores plot, it was possible to confirm the significant difference between the two levels of the factor "year": all samples collected in 2019 have negative scores, whereas almost all the samples collected in 2020 have positive scores, highlighting the high difference between the two levels of this factor. To explain this difference, in Figure 4b the corresponding loadings plot is reported, where it can be observed that the year 2020 samples present higher contents of almost all the molecules investigated in the study, except for 2-hexenal and myrcene, which do not contribute to explain the difference between the two years.

Figure 5a,b shows the scores and loadings plots for the effect of factor "cut", respectively. They are represented in the same way as for the factor "year". In this case, the scores plot confirms that there is a significant difference between the second and fourth cuts, even if it is not as marked as for the other main factors. In particular, scores of samples from 10 to 18 (4th cut, year 2019) present both positive and negative values in an irregular pattern. From the loadings plot, it is possible to observe that samples collected at the fourth cut present mainly a higher content of myrcene, eugenol and linalool, with respect to the second cut samples. β-caryophyllene and 2-hexenal contribute in the same direction but to a lesser extent. A slightly lower content of estragole characterizes the second cut. In general, for the factor "cut", not all the samples characterized by the same conditions behave similarly, as the effect of "cut" is of the same entity of its interactions with year and variety, as highlighted in Table 4. However, the general trend suggests that the influence of this factor on basil's aromatic profile cannot be neglected.

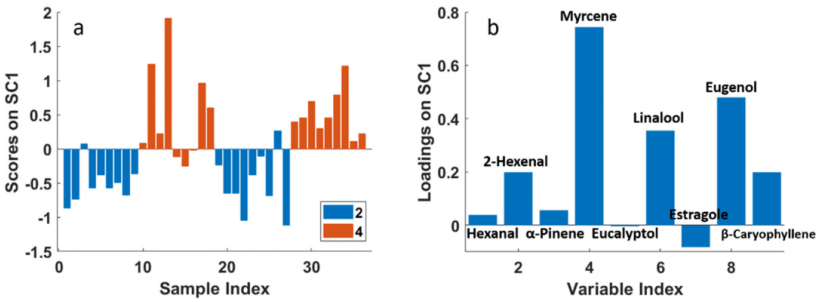


Figure 5. SCA of the effect matrix "cut". (**a**) Scores plot (SC1); (**b**) variable loadings (SC1).

Results of SCA for the factor "variety" are represented in Figure 6. In this case, since the factor "variety" was varied at three levels, two components (SCs) were necessary to describe its effect. The first SC clearly describes the difference between Var. 9 with respect to Var. 5 and Italiano Classico varieties. Var. 9 presented a higher content of almost all the molecules considered in this study, especially eucalyptol, estragole, and α-pinene, which gave a balsamic connotation to the odour (Table 2). On the other hand, the second SC shows the difference between Var. 5 and Italiano Classico varieties, less marked than the difference described by SC1. In this case, the compounds mainly responsible for this difference are hexanal and 2-hexenal, which are in greater quantity in the Italiano Classico variety, whereas Var. 5 is characterized by slightly higher quantities of eugenol, β-caryophyllene, α-pinene, estragole and eucalyptol.

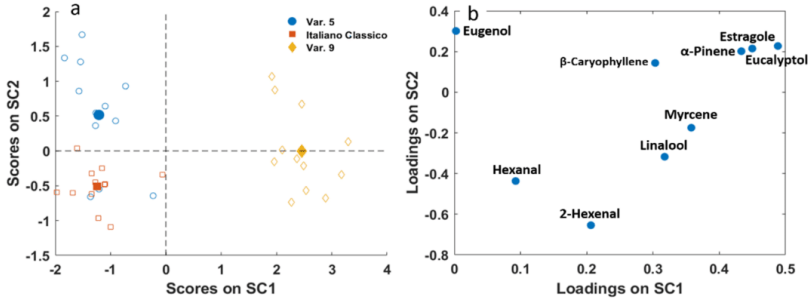


Figure 6. SCA of the effect matrix "variety". (**a**) SC1 vs. SC2 scores plot with projected residuals (empty symbols); (**b**) variable loadings (SC1 vs. SC2).

To deeply investigate the effect of considered factors on basil's aromatic profile, their second order interactions were also examined. Figure 7 shows the effect of the interaction between the factors "year" and "variety". It is possible to observe how Var. 9 is extremely different from the other two varieties, as it shows the opposite behaviour in SC1, i.e., Var. 9 samples collected in 2020 (negative SC1 values) have a higher content of almost all the considered molecules (negative SC1 loadings, except for 2-hexenal and hexanal close to zero) with respect to samples of the same variety collected in 2019. At variance, the other two varieties are richer in flavours in 2019 than in 2020. Italiano Classico and Var. 5 show the opposite behaviour with respect to year in SC2: the first is richer in flower/fruity aroma (higher myrcene and linalool) and lower in α-pinene and hexanal in 2019 with respect to 2020, and the opposite holds for Var. 5. Thus, it is worth noting how the variation of the factor "year" changes the chemical composition of samples of the same variety.

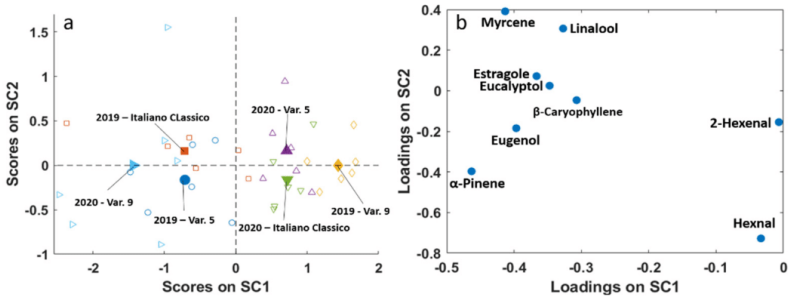


Figure 7. SCA of the effect matrix interaction "year x variety". (**a**) SC1 vs. SC2 scores plot with projected residuals (empty symbols); (**b**) variable loadings (SC1 vs. SC2).

The same pattern can be observed in Figure 8, which describes the effect of the interaction between the factors "cut" and "variety". In this case, the variation of factor "cut" is the one that strongly changes the chemical composition of samples characterized by the same variety, even if it does it to a lesser extent than the factor "year". High SC1 values correspond to a high 2-hexenal content, whereas low SC2 values are linked to high eugenol values.

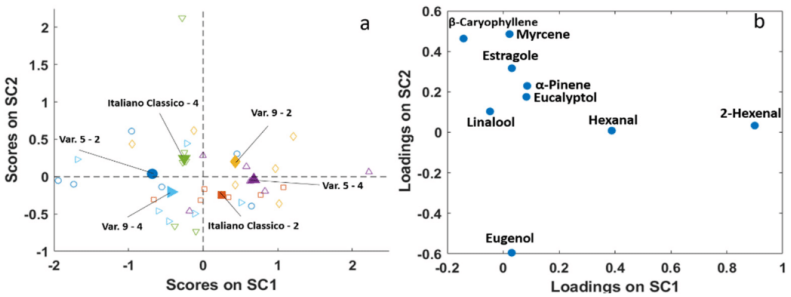


Figure 8. SCA of the effect matrix interaction "cut x variety". (**a**) SC1 vs. SC2 scores plot with projected residuals (empty symbols); (**b**) variable loadings (SC1 vs. SC2).

Considering the projected residuals, the differences are appreciable mainly in SC1, where Italiano Classico and Var. 9 show the same behaviour, being richer in floral/fruity flavours in cut 4 with respect to 2, while the opposite holds for Var. 5.

4. Conclusions

The results obtained support the use of a fast-GC based electronic nose for rapid assessment of basil aroma; in fact, the main molecules perceived as persistent by olfactometry (GC/O) are identifiable and quantifiable. In agreement with previous literature, it has been observed that the aroma composition is not only a distinctive trait of variety, but the content of each specific molecule varies with agronomic year and cut period. On the one hand, this renders more problematic the choice of a specific variety to be cultivated to achieve a desired flavor profile; on the other hand, it may help focus on the varieties showing more stability with respect to the agronomic variability. In terms of percentage of variance, the cut affects the aroma less with respect to year and variety. The effect of year seems to be a bulk effect affecting the content more than the type of molecules found in the aroma.

Author Contributions: Conceptualization, M.C., C.D. and A.D.; methodology, M.C., A.D., D.B., C.D. and L.S.; software, C.D. and L.S.; validation, M.C., A.D. and C.D.; investigation, A.D., C.D. and L.S.; resources, A.D.; data curation, A.D., C.D. and L.S.; writing—original draft preparation, A.D., C.D. and L.S.; writing—review and editing, M.C., A.D., C.D. and L.S.; supervision, M.C. and C.D.; project administration, A.D.; funding acquisition, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available on request from the authors.

Acknowledgments: The authors acknowledge Flavio Bertinaria for supplying basil samples and Federica Quaini for the sensorial descriptions.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Not available.

References

1. Qing, X.L.; Chiou, L.C. Basil (*Ocimum basilicum* L.) Oils. In *Essential Oils in Food Preservation, Flavor and Safety*; Preedy, V.R., Ed.; Academic Press: Cambridge, MA, USA, 2016; pp. 231–238.
2. Paton, A.; Harley, R.M.; Harley, M.M. *Ocimum*—An overview of relationships and classification. In *Medicinal and Aromatic Plants—Industrial Profiles*; Holm, Y., Hiltunen, R., Eds.; Harwood Academic: Amsterdam, The Netherlands, 1999; pp. 1–38.
3. Eileen, M.K.; Emily, D.N. Variations in phenolic composition and antioxidant properties among 15 basil (*Ocimum basilicum* L.) cultivars. *Food Chem.* **2011**, *128*, 1044–1050.
4. Grayer, J.R.; Kite, G.C.; Goldstone, F.J.; Bryan, S.E.; Paton, A.; Putievsky, E. Infraspecific taxonomy and essential oil chemotypes in sweet basil, Ocimum basilicum. *Phytochemistry* **1996**, *43*, 1033–1039. [CrossRef]
5. De Masi, L.; Siviero, P.; Esposito, C.; Castaldo, D.; Siano, F.; Laratta, B. Assessment of agronomic, chemical and genetic variability in common basil (*Ocimum basilicum* L.). *Eur. Food Res. Technol.* **2006**, *223*, 273–281. [CrossRef]
6. Salvadeo, P.; Boggia, R.; Evangelisti, F.; Zunin, P. Analysis of the volatile fraction of "Pesto Genovese" by headspace sorptive extraction (HSSE). *Food Chem.* **2007**, *105*, 1228–1235. [CrossRef]
7. Murarikova, A.; Tazky, A.; Neugebaureova, J.; Plankova, A.; Jampilek, J.; Mucaji, P.; Mikus, P. Characterization of Essential Oil Composition in Different Basil Species and Pot Cultures by a GC-MS Method. *Molecules* **2017**, *22*, 1221. [CrossRef] [PubMed]
8. Omer, E.A.; Said-Al, H.A.H.A.; Hendawy, S.F. Production, Chemical Composition and Volatile Oil of Different Basil Species/Varieties Cultivated under Egyptian Soil Salinity Conditions. *Res. J. Agric. Biol. Sci.* **2008**, *4*, 293–300.
9. Southwell, I.A.; Russel, M.F.; Davies, N.W. Detecting traces of methyl eugenol in essential oils: Tea tree oil, a case study. *Flavour Fragr. J.* **2011**, *26*, 336–340. [CrossRef]
10. Lee, S.J.; Umano, K.; Shibamoto, T.; Lee, K.G. Identification of volatile components in basil (*Ocimum basilicum* L.) and thyme leaves (*Thymus vulgaris* L.) and their antioxidant properties. *Food Chem.* **2005**, *91*, 131–137. [CrossRef]
11. Leonardos, G.; Kendall, D.; Barnard, N. Odor Threshold Determinations of 53 Odorant Chemicals. *J. Air Pollut. Control Assoc.* **1969**, *19*, 91–95. [CrossRef]
12. Plotto, A.; Margaria, C.A.; Goodner, K.L.; Baldwin, E.A. Odour and flavour threshold for key aroma components in an orange juice matrix: Terpenes and aldehydes. *Flavour Fragr. J.* **2004**, *19*, 491–498. [CrossRef]
13. Bertoli, A.; Lucchesini, M.; Mensuali-Sodi, A.; Leonardi, M.; Doveri, S.; Magnabosco, A.; Pistelli, L. Aroma characterization and UV elicitation of purple basil from different plant tissue cultures. *Food Chem.* **2013**, *141*, 776–787. [CrossRef]
14. Manzini, S.; Durante, C.; Baschieri, C.; Cocchi, M.; Marchetti, A.; Sighinolfi, S. Optimization of a Dynamic Headspace—Thermal Desorption—Gas Chromatography/Mass Spectrometry procedure for the determination of furfurals in vinegars. *Talanta* **2011**, *85*, 863–869. [CrossRef]
15. Biasioli, F.; Yeretzian, C.; Märk, T.D.; Dewulf, J.; Van Langenhove, H. Direct-injection mass spectrometry adds the time dimension to (B)VOC analysis. *Trends Anal. Chem.* **2011**, *30*, 1003–1017. [CrossRef]
16. Cocchi, M.; Durante, C.; Marchetti, A.; Armanino, C.; Casale, M. Characterization and discrimination of different aged 'Aceto Balsamico Tradizionale di Modena' products by head space mass spectrometry and chemometrics. *Anal. Chim. Acta* **2007**, *589*, 96–104. [CrossRef]
17. Capozzi, V.; Yener, S.; Khomenko, I.; Farneti, B.; Cappellin, L.; Gasperi, F.; Scampicchio, M.; Biasioli, F. PTR-ToF-MS Coupled with an Automated Sampling System and Tailored Data Analysis for Food Studies: Bioprocess Monitoring, Screening and Nose-space Analysis. *J. Vis. Exp. JoVE* **2017**, *123*, e54075. [CrossRef]

18. Lu, Y.; Gao, B.; Chen, P.; Charles, D.; Yu, L. Characterisation of organic and conventional sweet basil leaves using chromatographic and flow-injection mass spectrometric (FIMS) fingerprints combined with principal component analysis. *Food Chem.* **2014**, *154*, 262–268. [CrossRef]
19. Black, C.; Chevallier, O.P.; Elliott, C.T. The current and potential applications of Ambient Mass Spectrometry in detecting food fraud. *Trends Anal. Chem.* **2016**, *82*, 268–278. [CrossRef]
20. Lu, H.; Zhang, H.; Chingin, K.; Xiong, J.; Fang, X.; Chen, H. Ambient mass spectrometry for food science and industry. *Trends Anal. Chem.* **2018**, *107*, 99–115. [CrossRef]
21. Meng, X.; Zhai, Y.; Yuan, W.; Lv, Y.; Lv, Q.; Bai, H.; Niu, Z.; Xu, W.; Ma, Q. Ambient ionization coupled with a miniature mass spectrometer for rapid identification of unauthorized adulterants in food. *J. Food Compos. Anal.* **2020**, *85*, 103333. [CrossRef]
22. Giannoukos, K.; Giannoukos, S.; Lagogianni, C.; Tsitsigiannis, D.I.; Taylor, S. Analysis of volatile emissions from grape berries infected with Aspergillus carbonarius using hyphenated and portable mass spectrometry. *Sci. Rep.* **2020**, *10*, 21179. [CrossRef]
23. Torres, M.N.; Valdes, N.B.; Almirall, J.R. Comparison of portable and benchtop GC–MS coupled to capillary microextraction of volatiles (CMV) for the extraction and analysis of ignitable liquid residues. *Forensic Chem.* **2020**, *19*, 100240. [CrossRef]
24. Ketola, R.A.; Short, R.T.; Bell, R.J. 2.24—Membrane Inlets for Mass Spectrometry. In *Comprehensive Sampling and Sample Preparation*; Pawliszyn, J., Ed.; Academic Press: Cambridge, MA, USA, 2012; pp. 497–533.
25. Zlotek, U.; Mikulska, S.; Nagajek, M.; Swieca, M. The effect of different solvents and number of extraction steps on the polyphenol content and antioxidant capacity of basil leaves (*Ocimum basilicum* L.) extracts. *Saudi J. Biol. Sci.* **2016**, *23*, 628–633. [CrossRef]
26. Jordán, M.J.; Quílez, M.; Luna, M.C.; Bekhradi, F.; Sotomayor, J.A.; Sánchez-Gómez, P.; Gil, M.I. Influence of water stress and storage time on preservation of the fresh volatile profile of three basil genotypes. *Food Chem.* **2016**, *13*, 169–177. [CrossRef]
27. Fratianni, F.; Cefola, M.; Pace, B.; Cozzolino, R.; De Giulio, B.; Cozzolino, A.; d'Acierno, A.; Coppola, R.; Logrieco, A.F.; Nazzaro, F. Changes in visual quality, physiological and biochemical parameters assessed during the postharvest storage at chilling or non-chilling temperatures of three sweet basil (*Ocimum basilicum* L.) cultivars. *Food Chem.* **2017**, *229*, 752–760. [CrossRef] [PubMed]
28. Acree, T.E. GC/olfactometry GC with a sense of smell. *Anal. Chem.* **1997**, *69*, 170A–175A. [CrossRef]
29. Alpha-MOS. Available online: https://www.alpha-mos.com/heracles-smell-analysis (accessed on 20 April 2021).
30. Kostyra, E.; Król, K.; Knysak, D.; Piotrowska, A.; Żakowska-Biemans, S.; Latocha, P. Characteristics of volatile compounds and sensory properties of mixed organic juices based on kiwiberry fruits. *Appl. Sci.* **2021**, *11*, 529. [CrossRef]
31. Huang, L.; Liu, H.; Zhang, B.; Wu, D. Application of electronic nose with multivariate analysis and sensor selection for botanical origin identification and quality determination of honey. *Food Bioprocess Technol.* **2015**, *8*, 359–370. [CrossRef]
32. Wojtasik-Kalinowska, I.; Guzeka, D.; Gorska-Horczyczaka, E.; Głabska, D.; Brodowska, M.; Sun, D.; Wierzbick, A. Volatile compounds and fatty acids profile in Longissimus dorsi muscle from pigs fed with feed containing bioactive components. *LWT Food Sci. Technol.* **2016**, *67*, 112–117. [CrossRef]
33. Melucci, D.; Bendini, A.; Tesini, F.; Barbieri, S.; Zappi, A.; Vichi, S.; Conte, L.; Gallina-Toschi, T. Rapid direct analysis to discriminate geographic origin of extra virgin olive oils by flash gas chromatography electronic nose and chemometrics. *Food Chem.* **2016**, *204*, 263–273. [CrossRef]
34. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.; Lamers, R.J.A.; Van Der Greef, J.; Timmerman, M.E. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef]
35. Jansen, J.J.; Hoefsloot, H.C.; van der Greef, J.; Timmerman, M.E.; Westerhuis, J.A.; Smilde, A.K. ASCA: Analysis of multivariate data obtained from an experimental design. *J. Chemom. J. Chemom. Soc.* **2005**, *19*, 469–481. [CrossRef]
36. Rudnitskaya, A.; Rocha, S.M.; Legin, A.; Pereira, V.; Marques, J.C. Evaluation of the feasibility of the electronic tongue as a rapid analytical tool for wine age prediction and quantification of the organic acids and phenolic compounds. The case-study of Madeira wine. *Anal. Chim. Acta* **2010**, *662*, 82–89. [CrossRef] [PubMed]
37. Firmani, P.; Vitale, R.; Ruckebusch, C.; Marini, F. ANOVA-Simultaneous Component analysis modelling of low-level-fused spectroscopic data: A food chemistry case-study. *Anal. Chim. Acta* **2020**, *1125*, 308–314. [CrossRef]
38. Grassi, S.; Lyndgaard, C.B.; Rasmussen, M.A.; Amigo, J.M. Interval ANOVA simultaneous component analysis (i-ASCA) applied to spectroscopic data to study the effect of fundamental fermentation variables in beer fermentation metabolites. *Chemom. Intell. Lab. Syst.* **2017**, *163*, 86–93. [CrossRef]
39. Anderson, M.; Braak, C.T. Permutation tests for multi-factorial analysis of variance. *J. Stat. Comput. Simul.* **2003**, *73*, 85–113. [CrossRef]
40. De Luca, S.; De Filippis, M.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Marini, F. Characterization of the effects of different roasting conditions on coffee samples of different geographical origins by HPLC-DAD, NIR and chemometrics. *Microchem. J.* **2016**, *129*, 348–361. [CrossRef]
41. Raina, A.; Kumar, A. Chemical characterisation of basil germplasm for essential oil composition and chemotypes. *J. Essent. Oils Bear. Plants* **2017**, *20*, 1579–1586. [CrossRef]
42. Lawrence, B.M. A further examination of the variation of *Ocimum basilicum* L. In *Flavour and Fragrances, Proceedings of the 10th International Congress of Essential Oils, Fragrances and Flavors: A World Perspective, Washington, DC, USA, 16–20 November 1986*; Lawrence, B.M., Mookerjee, B.D., Willis, B.J., Eds.; Elsevier: Amsterdam, The Netherlands, 1988; pp. 161–170.

PAPER 2

chemosensors

MDPI

Article

Fast GC E-Nose and Chemometrics for the Rapid Assessment of Basil Aroma

Lorenzo Strani [1], Alessandro D'Alessandro [1,2], Daniele Ballestrieri [2], Caterina Durante [1,*] and Marina Cocchi [1]

[1] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy; lostrani@unimore.it (L.S.); alessandro.dalessandro@barilla.com (A.D.); marina.cocchi@unimore.it (M.C.)
[2] Barilla G. e R. Fratelli, Via Mantova 166, 43122 Parma, Italy; daniele.ballestrieri@barilla.com
* Correspondence: cdurante@unimore.it; Tel.: +39-059-2058-554

Abstract: The aim of this work is to assess the potentialities of the synergistic combination of an ultra-fast chromatography-based electronic nose as a fingerprinting technique and multivariate data analysis in the context of food quality control and to investigate the influence of some factors, i.e., basil variety, cut, and year of crop, in the final aroma of the samples. A low = level data fusion approach coupled with Principal Component Analysis (PCA) and ANOVA—Simultaneous Component Analysis (ASCA) was used in order to analyze the chromatographic signals acquired with two different columns (MXT-5 and MXT-1701). While the PCA analysis results highlighted the peculiarity of some basil varieties, differing either by a higher concentration of some of the detected chemical compounds or by the presence of different compounds, the ASCA analysis pointed out that variety and year are the most relevant effects, and also confirmed the results of previous investigations.

Keywords: basil; aroma; fast GC; electronic nose; untargeted fingerprint; PCA; ASCA; cut; variety

Citation: Strani, L.; D'Alessandro, A.; Ballestrieri, D.; Durante, C.; Cocchi, M. Fast GC E-Nose and Chemometrics for the Rapid Assessment of Basil Aroma. *Chemosensors* **2022**, *10*, 105. https://doi.org/10.3390/chemosensors10030105

Academic Editors: Larisa Lvova, Alisa Rudnitskaya and Federico Marini

Received: 7 February 2022
Accepted: 8 March 2022
Published: 10 March 2022

1. Introduction

Aromatic herbs of the *Laminacae* family are largely employed worldwide in culinary and health-related uses [1]. Among them, basil is largely used and very appreciated for its distinctive flavour, and its essential oils possess numerous health properties. Basil's origin dates back to over thousands of years; its name seems to derive from the ancient Greek "basilikon" (plant of the king), seemingly given for its peculiar characteristics [2]. There is a large number of basil cultivars and, for this reason, a standardized descriptor list, based on morphological characteristics, was developed by the International Union of Protection of New Varieties Plants (UPOV) [3]. In this list, *O. basilicum* is divided into six distinct morphotypes: 1. purple A, 2. purple B, 3. purple C, 4. lettuce, 5. small leaves, and 6. true basil [4]. Basil flavor is composed of different classes of molecules, such as ketones, alcohols, terpenoids, and esters [5], and for these reasons, a further classification scheme was proposed considering the different chemotypes: 1. high-linalool, 2. linalool/trans-α-bergamotene, 3. linalool/estragole, 4. linalool/trans-methyl cinnamate, and 5. high-estragole [6,7].

Basil has a relevant place in the Italian culinary culture, in the context of which it is largely used and appreciated [8]. There are different basil varieties [9] and are used, for example, in "pesto", a typical green sauce of the Italian region Liguria, in which a linalool/estragole basil chemotype prevails. In the last few years, in Italy, basil demand has increased: from 2015 to 2020, the harvested surface was more than doubled as was the produced quantity [10]. In this context, the selection of new varieties with improved agronomic characteristics and richer in appreciated flavor notes also became a relevant aspect. Traditionally, basil for food industry use has been cultivated on open fields, but greenhouses are sometimes used in early or late crops for productivity reasons. Normally,

in warmer climates, such as Italy, three-to-five cuts per harvesting year can be carried out; the first cut usually begins in late spring or early summer and the following cuts after about 20 days, depending on the weather conditions, and just before or at the start of flowering [11,12].

The development of 'artificial senses' for the evaluation of food quality and consumer preferences is nowadays well established [13]. In fact, on the one hand, they mimic food perceptions, and, on the other hand, they may furnish a quick evaluation and characterization of specific food attributes. In particular, under the general term electronic nose (e-nose) are comprised all types of sensors capable of detecting volatile organic compounds (VOCs), and include optical, electrical, electrochemical and mass-based detection [14,15]. Despite their different mechanisms, most of these sensors show non-specific recognition since they interact non-selectively with volatile molecules. In recent years, a new generation of e-nose instruments, based on ultra-fast gas chromatography with flame ionization detection (FID), i.e., fast GC-enose, has emerged as an appealing technology for VOC detection [16,17]. In fact, it shares the fast-screening capability of other types of e-noses, while allowing, at the same time, specificity and the putative identification of the detected molecules, which can be afterward confirmed by using a chromatographic run with standards or by GC-MS.

In order to characterize the basil flavor pattern, many analytical methods have been developed [18], such as the solvent extraction of the essential oil, or the direct sampling of the released volatile molecules by means of different analytical tools [19]. In fact, the more common tools are based on the direct collection of the headspace, or the trapping of the volatile molecules by Solid-Phase Micro Extraction (SPME) or by Head-Space Sorptive Extraction (HSSE) [20], while for the determination of the essential oil, gas chromatography (GC) is mainly employed, either coupled with mass spectrometry (MS) to have an identification, or just using flame ionization detection (FID), if identification is not the main concern.

In a previous study [21], our team developed an analytical method, based on e-nose ultrafast GC-FID, to characterize the basil flavor profile of some of the varieties currently employed in the production of Italian pesto sauce. Among the more than thirty peaks detected, only eighteen were tentatively identified on the basis of Kovats relative retention indices, and finally nine were confirmed by the analysis of the pure molecules. For these nine molecules, quantification was performed, constructing, for each one, a calibration curve with internal standards. These chemical markers allowed a partial chemical characterization of basil aroma profiles, and a differentiation of basil samples according to the studied agronomic factors.

The possibility to observe the complete chromatogram in an unsupervised way was the natural progression to fully benefit from the potential of the fast GC method. To this aim, in the present paper, the raw chromatographic signals, acquired in a very short time (110 s) by two different GC columns, are integrated according to a low-level data fusion approach [22,23], instead of considering (and quantifying) only the nine a priori known markers and the outcome of a single column. In addition, a higher number of basil samples collected from 2019 to 2021 (this year has not been previously considered) are measured, at the same time that the number of varieties studied is increased. Finally, the focus is the extraction of reliable chemical information from the raw signals aided by proper data analysis and preprocessing tools. In this way, without the need and the effort of the identification and quantification of specific markers, it is possible to study the different factors linked to production aspects and their influence on the product quality. This kind of approach could be easily and rapidly exported to other products where the knowledge of the individual molecules is more challenging or time consuming.

Multivariate data analysis pipeline included proper preprocessing, exploratory analysis by Principal Component Analysis (PCA), and ANOVA Simultaneous Component Analysis (ASCA) [24] to assess the effect of varieties, cuts period and harvesting years (2019, 2020 and 2021) on basil aroma. These are very critical aspects to consider when planning

the basil agronomic campaign in order to control the quality of pesto sauce, which is the product of interest.

2. Materials and Methods

2.1. Basil Plants

Plants of basil (*Ocimum basilicum*) of 20 commercial varieties of the "Genovese" type have been supplied by local producers over three different harvest years from 2019 to 2021. The varieties name was declared as code for confidentiality reasons and only the "Italiano Classico" variety was clearly indicated due to its largely commercial use. A total of 253 samples were collected and analysed.

Each basil variety was collected at different plant ages indicated as "cut". The plants were cut leaving about 5–6 cm from the soil, to allow the plants to regrow before the next cut. Typically, the first cut (labeled as 1 in Table 1) is performed after about 40 days from sowing and, then, the following cuts (numbered in time order from 2 to 5 in Table 1) after about 20 days each, depending on the weather and the agronomic conditions. Details of all samples (352 in total) are reported in Table 1. Varieties and cuts were not regularly varied during the three years because of company and producer constraints.

Table 1. Samples analyzed during the three years with the indication of the number of samples considered for each cut and, in italics, the number of replicates for each sample.

Harvesting Year	Basil Variety	Cut in Bold (n° of Samples; *Total Replicates*)				
2019	Italiano Classico	**1** (5; *18*)	**2** (2; *6*)	**3** (2; *6*)	**4** (2; *6*)	
	variety 5	**1** (1; *3*)				
	variety 7	**1** (2; *9*)				
	variety 9	**1** (1; *5*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 13	**1** (2; *3*)				
	variety 14		**2** (1; *3*)	**3** (1; *2*)	**4** (1; *3*)	
	variety 17	**1** (2; *5*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 18	**1** (2; *33*)				
	variety 19	**1** (2; *6*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
2020	Italiano Classico		**2** (2; *6*)	**3** (1; *3*)	**4** (2; *6*)	
	variety 1		**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 3		**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 5		**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 6				**4** (1; *3*)	
	variety 9				**4** (1; *3*)	
	variety 10			**3** (1; *3*)		
	variety 12		**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 14		**2** (1; *3*)		**4** (1; *3*)	
2021	Italiano Classico	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 2	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 4	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 8	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 9	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	**5** (1; *3*)
	variety 11	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 12	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 14	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	**5** (1; *3*)
	variety 15	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	
	variety 16	**1** (1; *3*)	**2** (1; *3*)	**3** (1; *3*)	**4** (1; *3*)	

2.2. Sample Preparation and VOC Sampling

Samples of the basil plants were collected in the early morning, typically from 4 to 8 a.m., and rapidly sent to the lab for characterization. All samples were analyzed within 6–8 h from the cut to minimize deterioration. For the analysis, about 30 g of the whole basil plant (leaves and stems), exactly weighted with a precision of 0.1 g, were hashed in a blender (Oster, Sunbeam Products Inc., Boca Raton, FL, USA) for 30 s in 300 mL of extraction solution at room temperature. The extraction solution was 100 g L^{-1} of NaCl and 6 mg kg^{-1} of ethyl iso-butyrate in water. NaCl was added to increase the volatile molecules release in the extraction headspace and ethyl iso-butyrate was added as an internal standard for the fast GC analysis. After the 30 s blending step, the suspension was left for 30 s, then 20 µL was collected and transferred in 20 mL amber vials that were immediately sealed and sent to analysis. Each extract was sampled at least three times in different vials. All reagents, standard and solvents were of analytical grade (Sigma Aldrich, St. Louis, MO, USA).

2.3. Heracles E-Nose Fast-GC Analysis

The analysis of the volatile molecules in the sample headspace was carried out using a Heracles II (Alpha MOS, Tuluse, France) ultra-fast chromatography electronic nose [25]. The e-nose consists of a double-columns ultra-fast-chromatography system, with FID detectors, interfaced with a PAL-RSI automatic headspace autosampler. Sample headspace air was collected and injected in the e-nose. The injected air was trapped on a Tenax TA polymer trap positioned before the columns. The two columns are mounted in parallel in the oven and have different polarities, MXT-5 (non-polar) and MXT-1701 (slightly polar); both have a length of 10 m, internal diameter of 0.18 mm and a phase thickness of 0.40 µm. A temperature ramp was employed, starting from 50 °C for 2 s, then increasing to 80 °C at 1 °C/s and finally reaching 250 °C at 3 °C/s. The total fast GC analysis time was 110 s. The carrier gas was hydrogen.

Each replicate of the extracted samples was loaded in the instrument auto sampler and incubated for 20 min at 40 °C before injection with 500 rpm agitation (5 s on, 2 s off). Then, 1 mL of air headspace was injected with a syringe at the temperature of 50 °C. Trap loading conditions were 18 s at 40 °C, then flashed to 250 °C for the release in the two columns at a split ratio 1:1.

The AlphaSoft v 16.0 software was used for a preliminary process of the data that were subsequently exported for further elaborations.

Volatile compounds were putatively identified on the basis of Kovats' relative retention indices (KI) and can be related to specific aromas that are collected in the AroChemBase v 7.0 database (Alpha MOS, Tuluse, France) built-in software. In this way, eighteen compounds were tentatively identified, as reported in a previous work [21].

2.4. Data Analysis

2.4.1. Data Preprocessing

Since the proper preprocessing of the different instrumental signals is very important to achieve trustworthy results, a preprocessing strategy was implemented to align the chromatograms.

The raw chromatograms, resulting from each of the two columns, were separately preprocessed as follows:

- First, they were normalized for the respective internal standard;
- Then, they were aligned by using the icoshift algorithm [26] applied by intervals, taking as reference the average signal. The intervals were manually defined, holding a single peak or small groups of peaks, as reported in Figure 1a. Alignment was necessary to compensate for the peaks shift, along retention time, among different chromatographic runs, which could introduce variability among samples not due to actual differences;

- The aligned chromatograms were baseline corrected by using the automatic weighted least squares algorithm (2nd order polynomial) [27];
- Considering that, in the analyzed chromatograms, the peaks' intensity and variance reflect the presence of major and minor constituents, it was important to use a procedure able to make the different chromatographic regions comparable in influence on the developed statistical models. In particular, block scaling to equal block variance (defining the blocks to be the same as the intervals used for the alignment with icoshift) was used, including column mean centering.

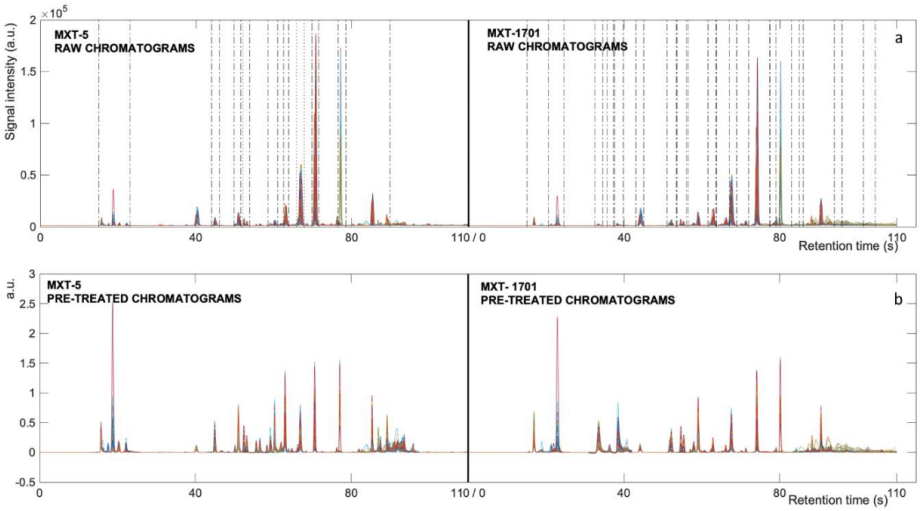


Figure 1. Collected chromatograms of basil samples (**a**) before and (**b**) after the different data pretreatments. (**a**) Dotted lines mark the limits of the different intervals used for the alignment and the scaling of the signals.

The preprocessed chromatograms are shown in Figure 1b.

A low-level data fusion approach was applied in order to simultaneously capture information coming from the analysis of samples through the two columns as well as to combine two potentially different sources of information. Indeed, from a chemical point of view, the slightly different polarity between the columns could highlight the presence of different analytes or obtain a better resolution, avoiding the loss of information due to possible co-elution issues. To this aim, the two singularly preprocessed chromatographic data sets were then concatenated in a single matrix of 352 (samples including replicates) × 20,002 (retention time points) dimensions. The MXT-5 and MXT-1701 chromatographic signals have a retention time ranging from 0 to 110 s sampled at 100 Hz, giving each 10,001 data points.

Prior to PCA, the concatenated data sets were block-scaled by considering as distinct data block each GC column (each data block comprises 10,001 variables, which are the respective sampled retention times), in order to let them equally contribute in PCA modelling.

2.4.2. ASCA

After data pretreatment (as detailed above in Section 2.4.1), the low-level fused chromatographic data (352 × 20,002 matrix dimensions) were subject to multivariate data analysis. As described in Table 1, samples varied according to three factors: harvesting year, variety and cut.

Principal Component Analysis (PCA) was applied on the entire data matrix to obtain a global overview of the trend, similarity and differences among the investigated samples according to the entire aroma profiles.

Furthermore, in order to assess the significance of the three factors (year, variety and cut) and their interactions, the ANOVA-Simultaneous Component Analysis (ASCA) method was used [24]. As a first step, ASCA performs an ANOVA, partitioning the data matrix X into the contribution of each factor or interaction, as shown in Equation (1):

$$X_c = X - 1m^T = X_1 + X_2 + X_3 + X_{1\times2} + X_{2\times3} + X_{1\times3} + X_{1\times2\times3} + X_{res} \quad (1)$$

where X_c is the centered data matrix, m^T is the mean profile of the samples, X (1, 2 and 3) is the main effect matrices, X (1 × 2, 2 × 3, 1 × 3 and 1 × 2 × 3) is the interaction effect matrices and X_{res} is the residuals matrix. Then, a Simultaneous Component Analysis (SCA) is performed, obtaining a scores matrix T and a loadings matrix P for each effect or interaction matrix, as described by Equation (2):

$$X_i = T_i P_i^T \quad (2)$$

ASCA needs balanced designs to provide reliable results. In order to avoid the construction of a design where the number of combinations per factor level is not equal, 18 conditions were selected from the original dataset for a total of 54 experiments, as shown in Table 2. Thus, in this model, three levels for factor "year" (2019, 2020 and 2021), three for factor "variety" (Italiano Classico, VAR 9 and VAR 14) and two levels for factor "cut" (2 and 4) were considered.

Table 2. Structure of the experimental design for ASCA for the years 2019, 2020 and 2021.

Year	Variety	Cut
2019	Italiano Classico	2
2019	Italiano Classico	4
2019	VAR 9	2
2019	VAR 9	4
2019	VAR 14	2
2019	VAR 14	4
2020	Italiano Classico	2
2020	Italiano Classico	4
2020	VAR 9	2
2020	VAR 9	4
2020	VAR 14	2
2020	VAR 14	4
2021	Italiano Classico	2
2021	Italiano Classico	4
2021	VAR 9	2
2021	VAR 9	4
2021	VAR 14	2
2021	VAR 14	4

Moreover, in order to further investigate the influence of varieties and cuts on basil aromatic profiles, another ASCA model was computed considering the year 2021 (where a higher number of varieties was cultivated), giving the sub-set of experiments described in Table 3. In this case, 9 basil varieties and 3 different cuts were inspected, for a total of 27 conditions and 81 experiments. It was not possible to investigate all levels for each

experimental factor, due to the limited varieties available that could be cultivated by a single producer.

Table 3. Structure of the experimental design for ASCA for the year 2021.

Variety	Cut
Italiano Classico	1
Italiano Classico	2
Italiano Classico	4
VAR 2	1
VAR 2	2
VAR 2	4
VAR 4	1
VAR 4	2
VAR 4	4
VAR 8	1
VAR 8	2
VAR 8	4
VAR 9	1
VAR 9	2
VAR 9	4
VAR 12	1
VAR 12	2
VAR 12	4
VAR 14	1
VAR 14	2
VAR 14	4
VAR 15	1
VAR 15	2
VAR 15	4
VAR 16	1
VAR 16	2
VAR 16	4

The significance of the effect of each design factor or interaction was evaluated through permutation tests (1000 randomizations), which compared the experimental sum of squares of each effect matrix with its related distribution under the null hypothesis [28].

2.4.3. Software

The raw chromatograms were imported and processed under a MATLAB 2020a (The MathWorks, Inc., Natick, MA, USA) environment. Chromatogram alignment was performed by using the icoshift 3.0, freely available on www.models.kvl.dk (last access on 7 March 2021). PCA and preprocessing were performed by PLS-Toolbox v. 8.9 (Eigenvector Inc., Manson, WA, USA). ASCA was carried out by using routines developed and kindly made available by Dr. F. Marini, University of Roma La Sapienza (Italy).

3. Results and Discussions

3.1. PCA Exploratory Analysis

In this first exploratory analysis, the aim was to obtain a general overview of the variation of the aroma volatile fraction of basil samples. Punctual considerations of the influence of harvested year, variety and cut could not be conducted, since it was not possible to plain a systematic sampling beforehand, due to company and producer constrains. Three principal components were considered according to their explained variances (58%). In Figure 2, the PC1 vs. PC2 score plot is reported, representing the different basil samples with different symbols and color as function of harvesting year and basil variety (Figure 2a) or cut and basil variety (Figure 2b).

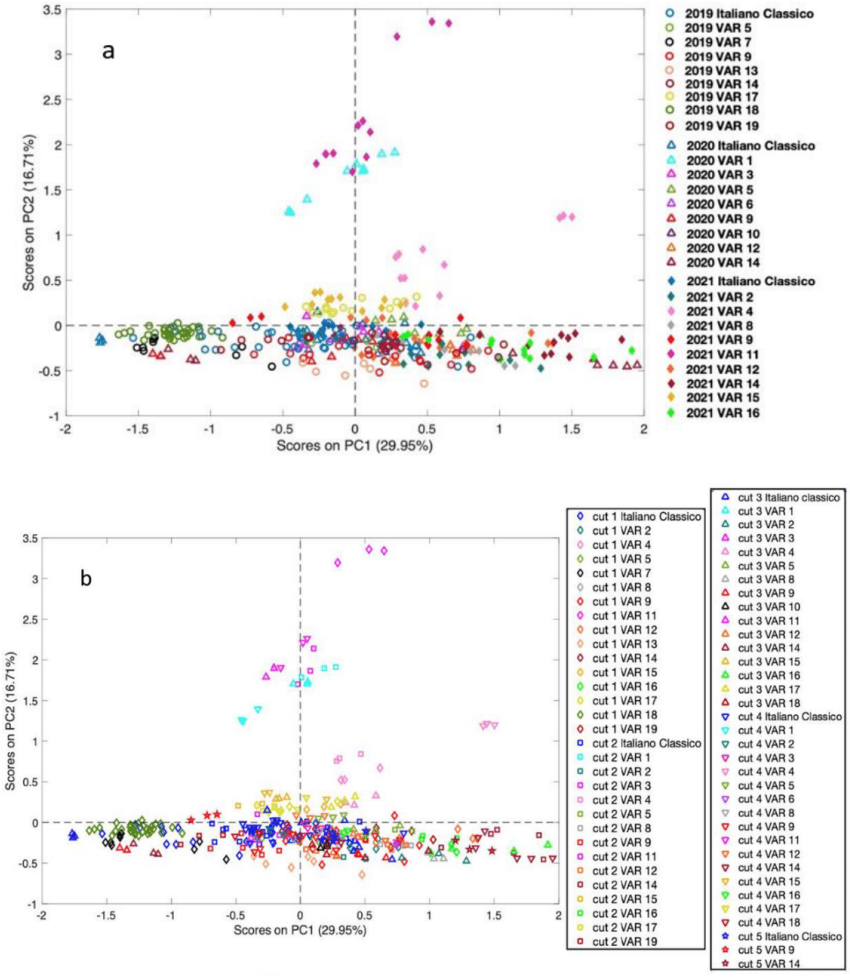


Figure 2. PC1 vs. PC2 score plots of basil samples. (**a**) Different symbols are used for each harvesting year (2019: circles; 2020: squares; 2021: triangles) and different colors for each basil variety. (**b**) Different symbols are used for each cut (first: diamonds; second: squares; third and fourth: upwards and downwards triangles, respectively; fifth: stars) and different colors for each basil variety.

From the score plot of the first two components, it is difficult to highlight a clear separation of samples according to all the different basil varieties, due to the slight differences in the flavor pattern among commercial varieties that belong to the same species (*O. basilicum*). However, interesting information can be pointed out. In particular, the VAR 1 (harvested only in 2019) and VAR 11 (harvested only in 2021) samples have the highest PC2 score values and leads to their separation from the other samples (Figure 2a). These varieties also present a trend, from higher to lower score values, according to their different cut (Figure 2b). Another peculiar variety seems to be VAR 4 (harvested only in 2021), with positive scores for both PC1 and PC2. This variety shows differences in aroma according to different basil cuts as well.

As far as the other samples are concerned, they are distributed along the first principal component, which seems to be the most responsible for the differences in the separation between the VAR 14 samples (higher positive PC1 score values) and first cut of VAR 7, VAR 18 and Italiano Classico (negative PC1 score values).

Furthermore, the in-depth analysis of the figure shows that two samples belonging to the third cut of VAR 16 (higher PC1 score values) seem to have quite a similar aroma profile to VAR 14.

No further observations to assess any pattern can be performed considering the different basil cuts, years and varieties, since it is not certain what the real cause is as some varieties were measured only in one year.

The score plot of the third component (Figure S1 reported in Supplementary Material) highlights the differences among the first basil cut of the VAR 8 and VAR 17 samples (higher positive score values) with respect to all the others.

From the PC1 loading plot (Figure 3a), for both MXT5 and MXT17 columns, it is possible to point out that, with almost all the loadings values being positive (from 40 to 110 s), the separation between the VAR 14 samples and the other basil varieties is mainly due to a global higher concentration of aroma compounds in these samples, and roughly speaking, most of the samples harvested in 2021 (positive PC1 score values) seem to present a similar trend.

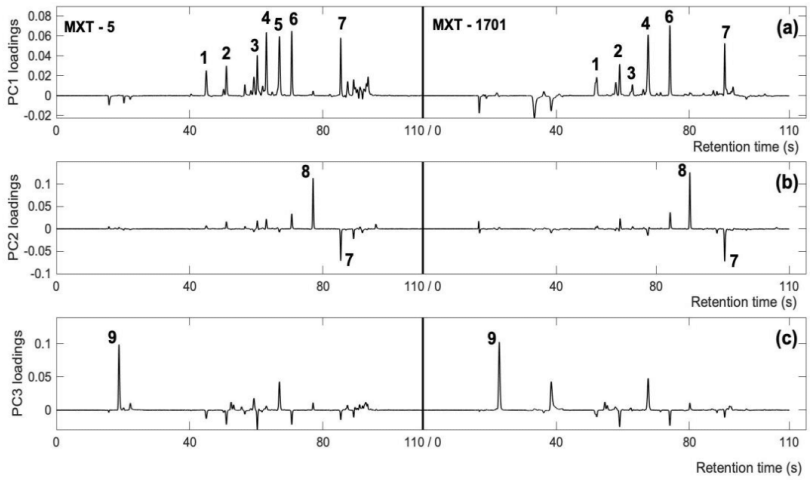


Figure 3. (**a**) PC1, (**b**) PC2 and (**c**) PC3 loading plots. Numbered peaks correspond to the volatile compounds putatively identified on the basis of Kovats's relative retention indices: (1) hexanal, (2) 2-hexanal, (3) 5-methylfurfural, (4) myrcene, (5) eucalyptol, (6) linalool, (7) β-caryophyllene, and (8) eugenol (9) not identified.

Notwithstanding the aim of the present study, which is to make a fast model to discriminate basil samples with an untargeted approach, some considerations on the presence of some chemical compounds can be presented on the basis of our previous study. Regarding the second principal component (Figure 3b), which is mainly responsible for the separation of VAR 1 and VAR 11 from the others, the same chromatographic regions (Rt, retention time: 76.8 s and 85.3 s for MXT-5 and 79.9 s and 90.4 s for MXT-17), for both the MXT-5 and MXT-17 columns, with the same trend (loadings value and sign), are relevant. Thus, both the estragole (Rt: 76.8 s and 79.9 s in MXT-5 and MXT-1701, respectively) and eugenol compounds (Rt: 85.3 s and 90.4 s in MXT-5 and MXT-1701, respectively), with high positive and negative loading values, respectively, are important to characterize VAR 1 and VAR 11. However, the samples belonging to these two varieties, presented a particular aroma, probably due to the presence of anethole, which co-elutes with estragole in both column separations.

As regards the third principal component (Figure 3c), unassigned compounds (in the first 40 s of both columns), which have positive loadings, seem more abundant in the VAR 8 and VAR 17 samples (located at positive scores values). Hence, further investigation will be conducted for the identification of these volatile compounds.

Notwithstanding the overall interpretation of PCA results, which offered some insights, more specific information is difficult to gain, since the contributions to variance of all the investigated factors (i.e., year, variety and cut) overlap. Therefore, after this preliminary investigation, the ASCA methodology was used in order to systematically assess the influence of each factor and their interaction on the basil aroma profile.

3.2. ASCA Results

The first ASCA model was computed according to the experimental design scheme shown in Table 2 (Section 2.4.2). The original data matrix variation was split in eight submatrices: three corresponding to the main effect of each experimental factor, three accounting for the effect of each second-order interaction, one describing the effect of the third-order interaction and one holding the residuals. The significance of all these effects was assessed by performing a permutation test, whose results are shown in Table 4. The *p*-value of all the inspected factors and interactions was lower than 0.001. However, the factors "variety" and "year" explained most of the data variance (39.9% and 24.8%, respectively), suggesting their higher influence on the aromatic profile of basil compared to the factor "cut". This can also be observed by the fact that explained variance values of interactions including "cut" are systematically lower than values related to interactions in which "cut" is not involved. Additionally, the third-order interaction effect explains less than 3% variance.

Table 4. Explained variance and *p*-values for main factors and their second and third order interactions.

Factor	Explained Variance (%)	*p*
Variety	39.9	<0.001
Year	24.8	<0.001
Year x Variety	8.5	<0.001
Year x Cut	7.2	<0.001
Cut	2.9	<0.001
Variety x Cut	2.5	<0.001
Year x Variety x cut	2.8	<0.001

Afterwards, the ASCA algorithm performed a SCA on each effect matrix individually, with the aim of interpreting the observed variation.

Figure 4a shows the score plot for the factor "year". The first component (SC1), which explains 67.7% of the total variance, describes the difference between the samples

harvested in 2019 and the samples harvested in 2020 and 2021. The loadings plot of the first component, shown in Figure 4b, explains this difference. In fact, the 2020 and 2021 samples appear to have a richer aroma profile, as the concentration of the compounds between 40 and 110 s, associated with statistically significant loadings, are higher compared to 2019 samples. On the other hand, 2019 samples present higher concentrations of unassigned peaks before 40 s mainly highlighted by the MXT-1701 column, confirming the need of further investigation for their identification.

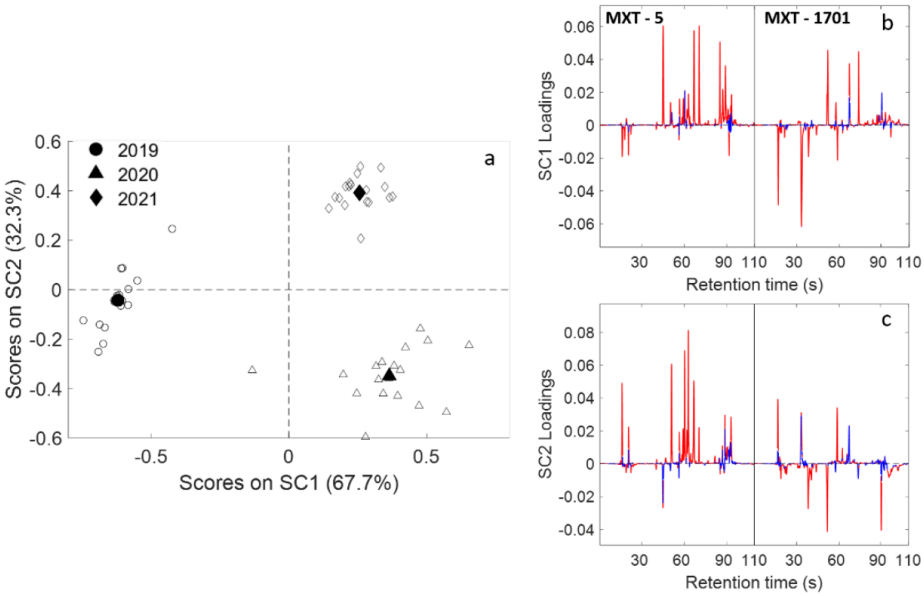


Figure 4. SCA for the effect of the factor "year". (**a**) SC1 vs. SC2 score plot. Empty symbols represent the projected residuals; (**b**) SC1 and (**c**) SC2 loadings plot. In the loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated with loadings statistically indistinguishable from zero.

The second component (SC2) and the related loadings plot (Figure 4c) show how the 2021 samples (positive scores values) present lower peaks in MXT-1701 that can be ascribed to 2-hexanal and β-caryophyllene (negative loadings values), but higher peaks assigned to all other compounds.

Figure 5a shows the score plot for the factor "variety". It can be observed that most of the explained variance (96.3%) describes how VAR 14 is different compared to Italiano Classico and VAR 9. Indeed, as shown by the loadings plot in Figure 5b, VAR 14 presents higher concentrations of all the chromatographic peaks, suggesting a richer aroma profile with respect to the other two varieties. SC2, even though the related explained variance is very low (3.7%), mainly shows how VAR 9 has more β-caryophyllene than Italiano Classico (Figure 5c), as their peaks are basically the only ones that had statistically significant results.

The results of the SCA for the effect of the interaction "year x variety" are reported in Figure 6. In the score plot (Figure 6a), it can be observed that SC1 describes the difference among VAR 14 samples throughout the years. In detail, the VAR 14 samples collected in 2020 presented a higher concentration of all aroma compounds compared to the ones collected in 2019 and 2021, as assumed by the loadings plot shown in Figure 6b. As regards Italiano Classico, the best year in terms of intensity of aroma profile is 2019, whereas for VAR 9, the years 2019 and 2021 were better than 2020.

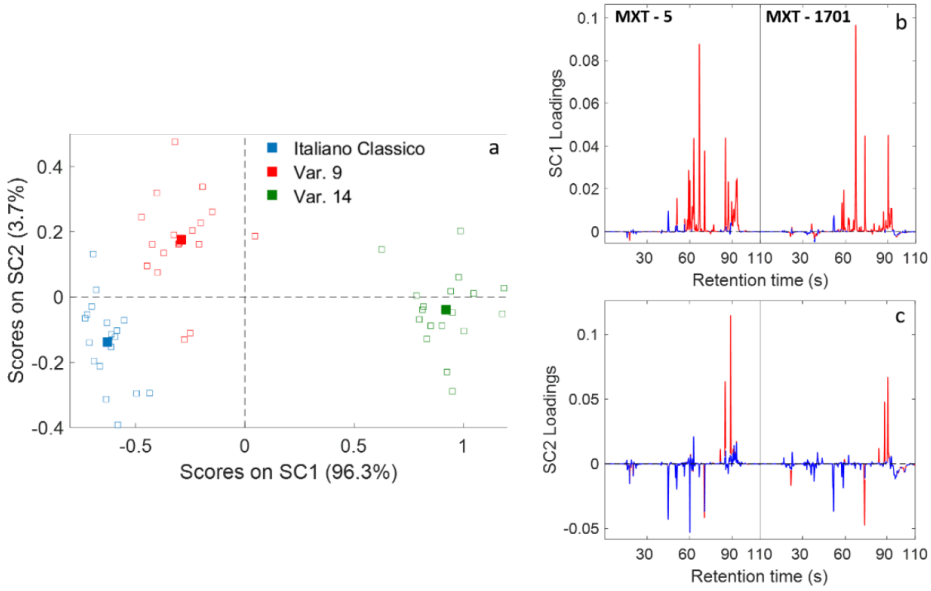


Figure 5. SCA for the effect of the factor "variety". (**a**) SC1 vs. SC2 score plot. Empty symbols represent the projected residuals; (**b**) SC1 and (**c**) SC2 loadings plot. In the loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated with loadings statistically indistinguishable from zero.

It can also be observed how VAR 14 appears to change more over time, having a higher variation through the years than the other two varieties.

Moreover, Italiano Classico is the basil variety that presents the lowest variability among its replicates. In fact, red and green samples in the score plot (VAR 9 and VAR 14, respectively) are more spread and farther apart, especially along SC2. This limits further comments about the difference between the years 2020 and 2021 with respect to the Italiano Classico samples (blue triangles and diamonds in Figure 6a, respectively), which is due to the statistically significant peaks between 50 and 70 s, linked to the majority of the aromatic compounds.

Regarding the factor "cut", the SCA showed how samples collected during cut 2 detain a richer aroma profile than samples acquired during cut 4. However, according to the authors, since this factor explained less than 3% of the total variance, these results are not relevant compared to the ones described above. Both for this reason and for the sake of brevity, plots related to the factor "cut" were not shown.

The second ASCA model was computed taking into account only samples collected in 2021. In this case, it was possible to build a balanced design, including nine varieties and three cuts, according to the scheme shown in Table 3 (Section 2.4.2). The data matrix was partitioned in four submatrices: two corresponding to the main effect of each experimental factor, one describing the effect of the second-order interactions and the residuals matrix. The results of the permutation test for the significance of the effects are shown in Table 5. As for the first ASCA model, also in this case, all the factors and their interactions were significant ($p < 0.001$). Furthermore, the explained variance for the factor "cut" (6.9%) was significantly lower than the variance explained by the factor "variety" (63.5%), suggesting, once again, the small impact of plant age on the basil aroma profile.

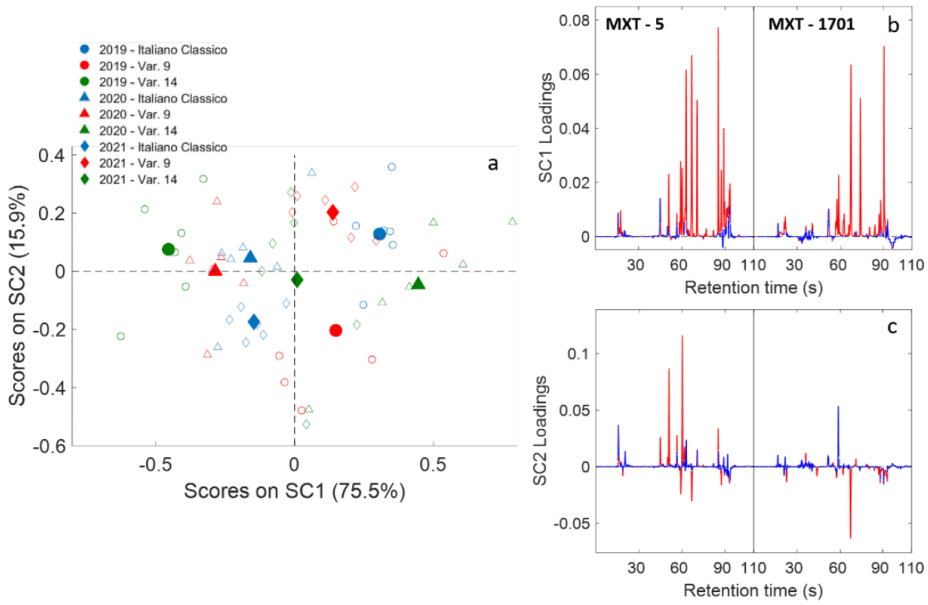


Figure 6. SCA for the effect of the interaction "year x variety". (**a**) SC1 vs. SC2 Score plot. Empty symbols represent the projected residuals; (**b**) SC1 and (**c**) SC2 loadings plot. In (**a**), the different colors refer to the different varieties (blue—Italiano Classico; red—VAR 9; green—VAR 14), whereas different symbols refer to different harvesting years (circles—2019; triangles—2020; diamonds—2021). In loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated with loadings statistically indistinguishable from zero.

Table 5. Explained variance and *p*-values for main factors and their second order interactions related to the ASCA model.

Factor	Explained Variance (%)	*p*
Variety	63.5	<0.001
Variety x Cut	20.3	<0.001
Cut	6.9	<0.001

The results related to the SCA on the "variety" effect matrix are shown in Figure 7.

From the score plot (Figure 7a), it is clear how the first principal component shows the difference between VAR 4 and all the other varieties. In the loadings plot (Figure 7b), it is shown that the peak that is mainly responsible for this difference can be ascribed to myrcene, of which VAR 4 is particularly rich. Observing SC2 scores and loadings (Figure 7c), it can be concluded that VAR 14 and VAR 16 present the richest aroma profiles, whereas Italiano Classico and VAR 15 have the poorest profiles.

Figure 8a shows the frequency histogram of the SC1 scores values for the different levels of the factor "cut". Eucalyptol and β-caryophyllene are less present in cut 4 samples, and in general, they are the compounds responsible for describing the difference between cut 4 samples and cut 1 and 2 samples, as shown in Figure 8b.

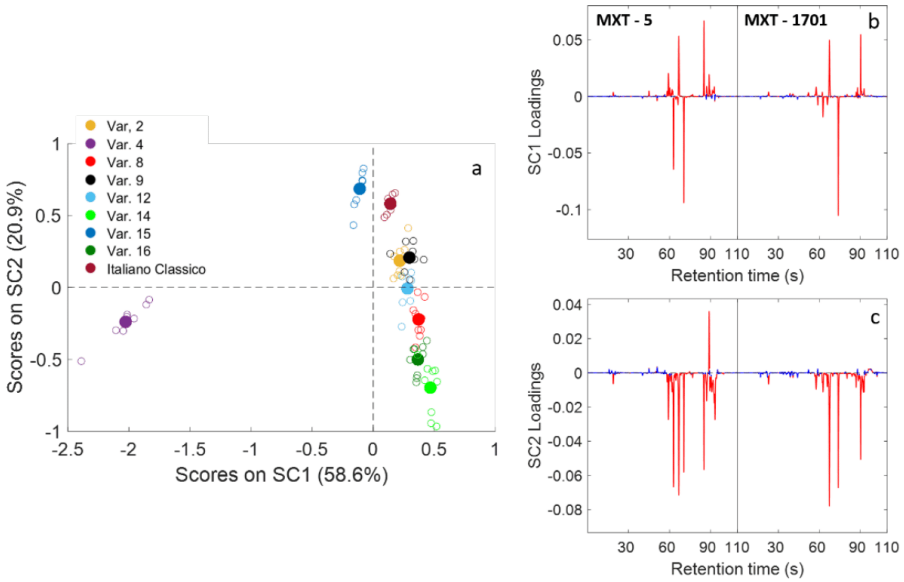


Figure 7. Results of ASCA performed on the 2021 samples. The SCA for the effect of the factor "variety". (**a**) SC1 vs. SC2 score plot. Empty symbols represent the projected residuals; (**b**) SC1 and (**c**) SC2 loadings plot. In loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated with loadings statistically indistinguishable from zero.

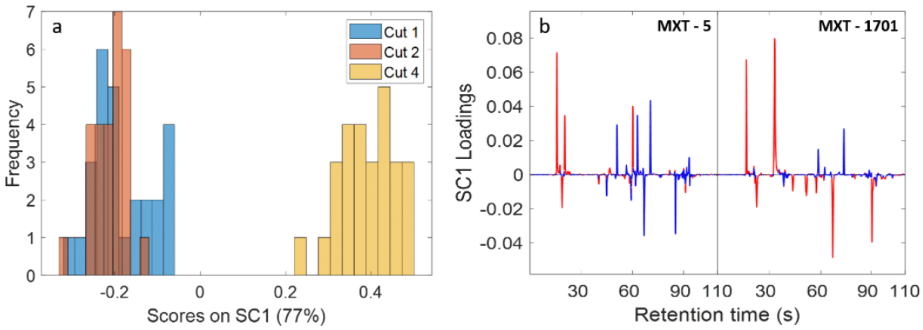


Figure 8. Results of ASCA performed on 2021 samples. The SCA for the effect of the factor "cut". (**a**) histograms of ASCA score frequency (with projected residuals) on SC1 for the different levels of factor "cut"; (**b**) SC1 Loadings plot. In loading plots, red lines indicate statistically significant regions, whereas blue lines indicate regions associated with loadings statistically indistinguishable from zero.

The ASCA results show how the entire aromatic profile has a significant influence in the discrimination of samples according to the investigated factors (i.e., years, variety and cut), highlighting the presence of new potential biomarkers (for instance the species with retention time in the first 30 s of the chromatogram or the ones falling in the area between the retention of 2-hexanal and 5-methylfurfural), which have not been quantified in this study, but that could be relevant in further investigations. For the sake of clarity, an example signal fingerprint with all the chemical analytes, putatively identified for both the chromatographic separations, is reported in Figure 9.

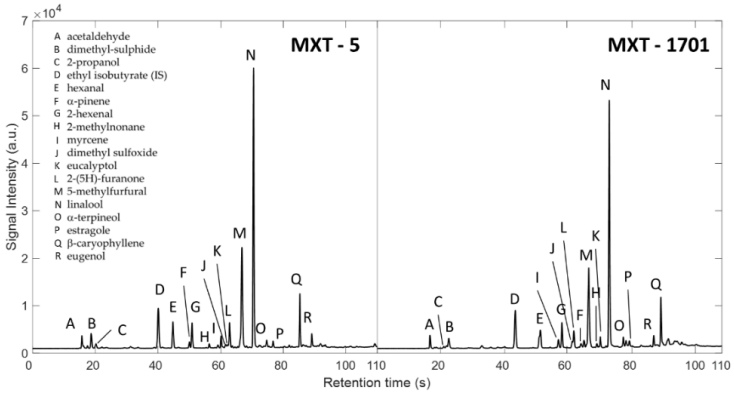


Figure 9. Chromatograms of the Italiano Classico variety obtained by elution on columns MXT-5 and MXT-1701 of Heracles II.

4. Conclusions

In this study, the development of a fast analytical screening strategy based on an ultra-fast chromatography e-nose and multivariate analysis was proposed as a useful tool for quality control of food. The proposed approach, relying on the simultaneous analysis of the chromatographic profiles coming from two GC-columns of different polarity, permits to explore fully the volatile profile of foodstuff and may represent a fast and simpler alternative to other chromatographic techniques. The chemical identification and quantification of the single chemical species, responsible for differentiation of the studied food products, can be undertaken on a few samples for a second time. In fact, once the main chromatographic peaks, mostly responsible for the differentiation between samples, have been underlined, their respective chemical species can be identified with a considerable reduction in costs and analysis time.

In particular, this approach was applied on the analysis of the basil samples involved in the production of Italian pesto sauce, where the entire e-nose signals, coming from two columns with different polarity, were fused and used as a fingerprint of the aroma profile. The obtained results highlighted the possibility of differentiating basil samples on the basis of the three investigated factors, years, cut and variety, taking also into account the interactions among them. The low-level data fusion approach allowed the computing of a single ASCA model, which effectively pointed out the different significant peaks between the two columns taken into account, thus underlining that enhanced information may be gained.

The knowledge of the influence of the investigated factors on the quality of basil is very important, since it may allow a company to achieve useful information both to plan future campaign strategies for the acquisition of the raw materials and to improve the quality of the final pesto sauce.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/chemosensors10030105/s1, Figure S1: PC3 scores vs. n° of sample.

Author Contributions: Conceptualization, M.C., C.D. and A.D.; methodology, M.C., A.D., D.B., C.D. and L.S.; software, C.D. and L.S.; validation, M.C., A.D. and C.D.; investigation, A.D., D.B., C.D. and L.S.; resources, A.D.; data curation, A.D., C.D. and L.S.; writing—original draft preparation, A.D., C.D., and L.S.; writing—review and editing, M.C., A.D., C.D. and L.S.; supervision, M.C. and C.D.; project administration, A.D.; funding acquisition, A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available on request from the authors.

Acknowledgments: The authors acknowledge Flavio Bertinaria for supplying basil samples and Federica Quaini for the sensorial descriptions.

Conflicts of Interest: Authors A.D. and D.B. are employed by Barilla G. e R. Fratelli SpA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflicts of interest.

References

1. Stefanaki, A.; van Andel, T. Mediterranean aromatic herbs and their culinary use. In *Aromatic Herbs in Food Bioactive Compounds, Processing and Applications*, 1st ed.; Galanakis, C.M., Ed.; Elsevier: Amsterdam, The Netherlands, 2021; pp. 93–121. [CrossRef]
2. Makri, O.; Kintzios, S. *Ocimum* sp. (Basil): Botany, Cultivation, Pharmaceutical Properties and Biotechnology. *J. Herbs Spices Med. Plants* **2008**, *13*, 123–150. [CrossRef]
3. UPOV International Union for the Protection of New Varieties of Plants, UPOV Code: OCIMUM_BAS TG/200/2 Date 2016:03-16. Available online: https://www.upov.int/edocs/tgdocs/en/tg200.pdf (accessed on 18 December 2021).
4. Hussain, A.I.; Anwar, F.; Sherazi, S.T.H.; Przybylski, R. Chemical composition, antioxidant and antimicrobial activities of basil (*Ocimum basilicum*) essential oils depend on seasonal variations. *Food Chem.* **2008**, *108*, 986–995. [CrossRef] [PubMed]
5. Anwar, F.; Alkharfy, K.M.; Mehmood, T.; Bakht, M.A.; Rehman, N.-U. Variation in chemical composition and effective antibacterial potential of ocimum basilicum L. essential oil harvested from different regions of Saudi Arabia. *Pharm. Chem. J.* **2021**, *55*, 187–193. [CrossRef]
6. Varga, F.; Carović-Stanko, K.; Ristić, M.; Grdiša, M.; Liber, Z.; Šatović, Z. Morphological and biochemical intraspecific characterization of *Ocimum basilicum* L. *Ind. Crops Prod.* **2017**, *109*, 611–618. [CrossRef]
7. da Silva, W.M.F.; Kringel, D.H.; de Souza, E.J.D.; da Rosa Zavareze, E.; Dias, A.R.G. Basil essential oil: Methods of extraction, chemical composition, biological activities, and food applications. *Food Bioprocess Technol.* **2021**, *15*, 1–27. [CrossRef]
8. Basil Leaves Market-Global Industrial Analysis, Size, Share, Trends, Growth and Forecast 201–1027. Available online: https://www.transparencymarketresearch.com/basil-leaves-market.html (accessed on 16 December 2021).
9. Barut, M.; Tansi, L.S.; Akyuz, A.M.; Karaman, S. Quality and yield of different basil (*Ocimum basilicum* L.) cultivars with various planting and cutting times under hot mediterranean climate. *Appl. Ecol. Environ. Res.* **2021**, *19*, 3115–3136. [CrossRef]
10. ISTAT Cultivar Data. Available online: http://dati.istat.it/Index.aspx?DataSetCode=DCSP_COLTIVAZIONI (accessed on 18 December 2021).
11. Davies, W.J.; Kozlowski, T.T. Stomatal responses to changes in light intensity as influenced by plant water stress. *For. Sci.* **1975**, *21*, 129–133.
12. De Martino, L.; Amato, G.; Caputo, L.; Nazzaro, F.; Scognamiglio, M.R.; De Feo, V. Variations in composition and bioactivity of ocimum basilicum cv 'Aroma 2' essential oils. *Ind. Crops Prod.* **2021**, *172*, 114068. [CrossRef]
13. Kiani, S.; Minaei, S.; Ghasemi-Varnamkhasti, M. Fusion of Artificial Senses as a Robust Approach to Food Quality Assessment. *J. Food Eng.* **2016**, *171*, 230–239. [CrossRef]
14. Calvini, R.; Pigani, L. Toward the Development of Combined Artificial Sensing Systems for Food Quality Evaluation: A Review on the Application of Data Fusion of Electronic Noses, Electronic Tongues and Electronic Eyes. *Sensors* **2022**, *22*, 577. [CrossRef] [PubMed]
15. John, A.T.; Murugappan, K.; Nisbet, D.R.; Tricoli, A. An Outlook of Recent Advances in Chemiresistive Sensor-Based Electronic Nose Systems for Food Quality and Environmental Monitoring. *Sensors* **2021**, *21*, 2271. [CrossRef] [PubMed]
16. Gorji-Chakespari, A.; Nikbakht, A.M.; Sefidkon, F.; Ghasemi-Varnamkhasti, M.; Valero, E.L. Classification of essential oil composition in Rosa damascena Mill genotypes using an electronic nose. *J. Appl. Res. Med. Aromat. Plants* **2017**, *4*, 27–34. [CrossRef]
17. Nie, J.Y.; Rong, L.; Zi-Tao, J.; Ying, W.; Jin, T.; Shu-Hua, T.; Yi, Z. Antioxidant activity screening and chemical constituents of the essential oil from rosemary by ultra-fast GC electronic nose coupled with chemical methodology. *J. Sci. Food Agric.* **2020**, *100*, 3481–3487. [CrossRef] [PubMed]
18. Sharma, S.; Kumari, A.; Dhatwalia, J.; Guleria, I.; Lal, S.; Upadhyay, N.; Kumar, A. Effect of solvents extraction on phytochemical profile and biological activities of two ocimum species: A comparative study. *J. Appl. Res. Med. Aromat. Plants* **2021**, *25*, 100348. [CrossRef]
19. Carovic-Stanko, K.; Ribic, A.; Grdisa, M.; Liber, Z.; Kolak, I.; Satovic, Z. In Identification and discrimination of *Ocimum basilicum* L. morphotypes. In Proceedings of the 46th Croatian and 6th International Symposium on Agriculture, Opatija, Croatia, 14–18 February 2011.
20. Klimánková, E.; Holadová, K.; Hajšlová, J.; Čajka, T.; Poustka, J.; Koudela, M. Aroma profiles of five basil (*Ocimum basilicum* L.) cultivars grown under conventional and organic conditions. *Food Chem.* **2007**, *107*, 464–472. [CrossRef]
21. D'Alessandro, A.; Ballestrieri, D.; Strani, L.; Cocchi, M.; Durante, C. Characterization of basil volatile fraction and study of its agronomic variation by asca. *Molecules* **2021**, *26*, 3842. [CrossRef] [PubMed]

22. Silvestri, M.; Bertacchini, L.; Durante, C.; Marchetti, A.; Salvatore, E.; Cocchi, M. Application of data fusion techniques to direct geographical traceability indicators. *Anal. Chim. Acta* **2013**, *769*, 1–9. [CrossRef] [PubMed]
23. Biancolillo, A.; Boquè, R.; Cocchi, M.; Marini, F. Data fusion strategy in food analysis. In *Data Handling in Science and Technology*; Cocchi, M., Ed.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 31, pp. 271–310.
24. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.; Lamers, R.J.A.; Van Der Greef, J.; Timmerman, M.E. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef] [PubMed]
25. Alpha-MOS. Available online: https://www.alpha-mos.com/heracles-smell-analysis (accessed on 20 January 2022).
26. Savorani, F.; Tomasi, G.; Engelsen, S.B. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson.* **2010**, *202*, 190–202. [CrossRef] [PubMed]
27. Wise, B.M.; Gallagher, N.B.; Bro, R.; Shaver, J.M.; Windig, W.; Koch, R.S. *Chemometrics Tutorial for PLS_Toolbox and Solo*; Eigenvector Research, Inc.: Wenatchee, WA, USA, 2006; pp. 173–174.
28. Anderson, M.; Braak, C.T. Permutation tests for multi-factorial analysis of variance. *J. Stat. Comput. Simul.* **2003**, *73*, 85–113. [CrossRef]

PAPER 3

foods

MDPI

Article

A Feasibility Study towards the On-Line Quality Assessment of Pesto Sauce Production by NIR and Chemometrics

Daniele Tanzilli [1,2], Alessandro D'Alessandro [1], Samuele Tamelli [1], Caterina Durante [1], Marina Cocchi [1,*] and Lorenzo Strani [1]

[1] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy; daniele.tanzilli@unimore.it (D.T.); alessandro.dalessandro@barilla.com (A.D.); samueletamelli1997@gmail.com (S.T.); caterina.durante@unimore.it (C.D.); lostrani@unimore.it (L.S.)
[2] Université de Lille, CNRS, LASIRE (UMR 8516), Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, 59000 Lille, France
* Correspondence: marina.cocchi@unimore.it

Abstract: The food industry needs tools to improve the efficiency of their production processes by minimizing waste, detecting timely potential process issues, as well as reducing the efforts and workforce devoted to laboratory analysis while, at the same time, maintaining high-quality standards of products. This can be achieved by developing on-line monitoring systems and models. The present work presents a feasibility study toward establishing the on-line monitoring of a pesto sauce production process by means of NIR spectroscopy and chemometric tools. The spectra of an intermediate product were acquired on-line and continuously by a NIR probe installed directly on the process line. Principal Component Analysis (PCA) was used both to perform an exploratory data analysis and to build Multivariate Statistical Process Control (MSPC) charts. Moreover, Partial Least Squares (PLS) regression was employed to compute real time prediction models for two different pesto quality parameters, namely, consistency and total lipids content. PCA highlighted some differences related to the origin of basil plants, the main pesto ingredient, such as plant age and supplier. MSPC charts were able to detect production stops/restarts. Finally, it was possible to obtain a rough estimation of the quality of some properties in the early production stage through PLS.

Keywords: MSPC charts; on-line; process monitoring; NIR; Basil; pesto production; PCA; PLS

Citation: Tanzilli, D.; D'Alessandro, A.; Tamelli, S.; Durante, C.; Cocchi, M.; Strani, L. A Feasibility Study towards the On-Line Quality Assessment of Pesto Sauce Production by NIR and Chemometrics. *Foods* **2023**, *12*, 1679. https://doi.org/10.3390/foods12081679

Academic Editors: Jordi Riu and Barbara Giussani

Received: 15 March 2023
Revised: 10 April 2023
Accepted: 14 April 2023
Published: 18 April 2023

1. Introduction

One of the most important aspects in food industrial production, in addition to basic safety and compliance requirements, is the capability to guarantee a constant quality of the final product, including all aspects from composition to appearance and taste. To achieve this aim, a lot of effort is spent monitoring the process, usually through univariate control charts and focusing most of the effort on monitoring the quality of the final product. However, operating in this way is not optimal when the food processing is complex, and production is massive. In fact, it may be difficult in this way to understand which are the Normal Operative Conditions (NOC) of the process. Since many parameters can change simultaneously and can be correlated, it is not easy for plant operators to detect the problem in a fast way in case anomalies or deviations occur [1]. In addition, although reference analyses are reliable and efficient in assessing the final product's quality, they provide slow responses as the sample must be collected, brought into the laboratory, and analyzed, being, at the same time, expensive in terms of money, operators' effort, and waste. For this reason, different types of sensors that can provide timely information are becoming more and more used for on-line quality probing from raw materials to the semi-finished and final products. It has been widely demonstrated that NIR spectroscopy has a powerful potential in monitoring food production processes [2–11], due to its ability to detect both chemical and physical changes in the samples. To cite a few applications: NIR has been used for

process monitoring in the dairy industry, from the prediction of raw milk composition to milk coagulation in cheese production and yogurt fermentation [11]; the fermentation processes in the wine and brewery industries; and the powdered ingredients mixing stage in different food matrices [10]. Thus, the on-line implementation of a NIR monitoring system is desired for several reasons: the timely handling of any possible faults, reducing products out of specification, thus reducing waste and economical loss. Moreover, if in addition to the data coming from process sensors controlling the machinery settings (such as the temperature, mixing rate, pressure, etc.) fused with NIR, it could become feasible to achieving a better understanding of processes, which could aid in designing more efficient and environmentally friendly processes [12,13]. However, it is still not so common in food production to have implemented systems for the data storage of retrieval process sensors. Nonetheless, companies are becoming increasingly interested in developing models that can achieve real-time monitoring and improve industrial processes.

However, it is difficult to handle, fuse, and interpret sensor data, as it is not possible to rapidly extract useful information from spectra and images without proper statistical tools. Thus, developing multivariate control charts based on latent variables and real-time prediction models, benefitting from the chemometric development in this area, is starting to be a recognized advantage in the industry.

It has been extensively demonstrated how Multivariate Statistical Process Monitoring/Control based on Latent Variables (MSPC-LVs) can lead to an efficient process monitoring [14–22].

The present work concerns a feasibility study to set up a model for the on-line monitoring of the pesto production process in the company Barilla, where, at the moment, a vision system (RGB camera) is monitoring the main raw material, i.e., basil, and a NIR probe installed in-line is monitoring the initial semi-finished product. The main aim of this preliminary feasibility study is the evaluation of the possible advantages that MSPC-LVs based on in-line acquired data can furnish both in terms of the possibility of estimating the quality of the finite product in real time and capturing the process evolution and the eventual departure from NOC. In this context, PCA models have been used to explore the data structure and the information they furnish. Furthermore, multivariate control charts for process monitoring based on NOC data were built. Lastly, a first attempt to obtain predictive models for the real-time prediction of main pesto quality parameters has been also carried out.

The focus has been on discussing the steps that were more critical for the models' development. Although the results are very preliminary, some interesting indications and directions for improvement could be formulated.

2. Materials and Methods

2.1. Process Description

The analyzed data were collected from the pesto sauce line during the 2020 harvesting season in a production plant owned by the company Barilla G. e R. Fratelli S.p.A., located near Parma, Italy. In this campaign, two different varieties of basil (*Ocimum basilicum*), the main ingredient of the sauce, have been provided by five local suppliers and continuously delivered to the process line. Each basil variety was harvested four times at different plant ages: the first cut was performed at 40 days, whereas the successive cuts were each carried out every 20 days.

At the beginning of the process line, a vision system (RGB camera) was installed that acquired images of basil plants while passing on the conveyor belt. The system was set to deliver some parameters in real time, such as the average and standard deviation values (every 15 s) of the R, G, and B channels and a rough estimation of the basil leaves' area in the acquired image (not always available at the same time intervals); however, the raw images were not always stored. Thus, in this work, only the R, G, and B parameters could be considered.

After this step, basil was mixed with salt and oil, forming an intermediate product, which was monitored on-line by a NIR probe. Then, all the other ingredients of the sauce were added to the intermediate product to complete the production and obtain the final product, whose quality was assessed by off-line laboratory analyses. A schematic representation of the process is reported in Figure 1. A critical issue when modeling on-line data for a continuous process is to establish the process timeline to match the sensors data acquired at different time steps with the same material; in other words, the considered variables should refer to the same sample to assemble a row of the data matrix. In this case, this step revealed particularly challenging, since the mixing of the intermediate product with the other ingredients (taking place after the NIR probe) was achieved in three distinct mixers that were emptied, transferring the crude pesto to the following processing steps sequentially, ensuring a continuous material flux. Thus, the residence time was established with the experts at the plant in order to correctly match the NIR spectra, corresponding to the intermediate material with the finished pesto at the end of the line, on which the quality parameters were acquired.

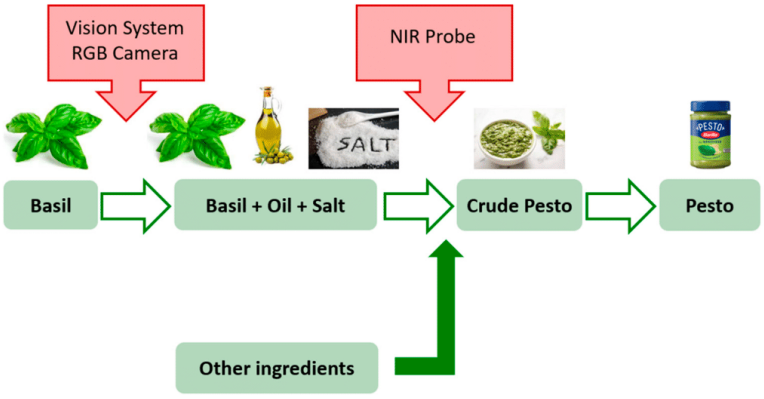


Figure 1. Schematic representation of pesto sauce production process.

In this study, data collected from May to August 2020 were analyzed, but not all the data recorded during this period were considered for model building, due to production pauses, instrument maintenance, and unreliable acquisitions. Finally, 459 data points were considered.

This is a second critical issue when assembling the data matrix since interruptions could be quite frequent. Since the on-line RGB continue to acquire the image of the same basil when a stop occurs at the raw material conveyor belt, an inspection of the RGB parameters' time trends with the identification of constant values as the indication of the stopping period was used. Moreover, the activation of the pump transferring the intermediate to the NIR probe was registered and was also used as an indication of stopping periods.

Finally, a data cleaning based on anomalous RGB values of the spectra was also accomplished.

2.2. *Reference Analysis*

Consistency parameters and lipids content have been considered for the assessment of pesto sauce quality. These parameters were assessed off-line by collecting pesto samples right after their production was complete.

The Consistency of pesto is evaluated measuring the flow of a standard volume of sample (100 cm^3) under its own weight. The flow could be related to the sample viscosity. To perform the measure, a Bostwick consistometer was used (ASTM F1080-93). This is a

stainless-steel slide with a reservoir of 5 × 5 × 4 cm, a mobile gate, two adjusting screws for planarity, and a track with ruler markings. The sample, conditioned to the temperature of 20 °C, was loaded into the reservoir. Then, the gate was opened, the timer was started, and sample flowed on the track. The consistency of the pesto was evaluated, measuring the distance in centimeters flowed in 30 s. Before the measure, the dedicated adjusting screws leveled the consistometer.

The total lipids content was determined by solvent extraction on a weighed sample aliquot (5 to 10 g). The extraction was conducted with an ethyl ether in a Soxhlet apparatus for 4 h. The sample was placed in a rotary evaporator and placed in an oven at 105 °C for 2 h to remove the solvent. The fat extracted was weighed at room temperature, and its content was expressed in a percentage, divided by the initial weight of the sample.

2.3. On-Line Instrumentation

A Sensure prototype camera (Sensure, Bergamo, Italy) was installed above the conveyor belt right after the basil plants were supplied, acquiring RGB images every 15 s. R, G, and B values were extracted by images and treated as separate variables.

A ProFoss spectrometer (Foss, Hillerød, Denmark) was used to collect the spectra of the basil, salt and oil mixture, namely, the intermediate product. The instrument was equipped with an optical fiber, whose probe was installed at the acquisition site on the process pipe. The spectra were acquired over the 1100–1650 nm spectral range in the transmission mode, with a nominal resolution of 0–5 nm and 64 scans per sample.

2.4. Data Analysis

The data analysis objectives were twofold: on the one hand, we evaluated the potentiality of establishing an on-line monitoring model (Section 2.4.3: Multivariate Control Charts; the results discussed in Section 3.2) capable of describing the natural variability inherent to the process and of capturing any eventual anomalous fluctuation, and, on the other hand, we aimed at establishing predictive models (Section 2.4.4: PLS Regression; the results discussed in Section 3.3) to evaluate the feasibility of the prediction of quality properties of the pesto sauce in real time.

However, prior to the model building, multivariate data exploration (Section 2.4.2: Principal Component Analysis; the results discussed in Section 3.1) has been a mandatory step to inspect the data structure and presence of deviating samples and to establish the time points corresponding to the normal operating conditions for the plant.

To ease readability, the applied preprocessing has been enclosed and detailed in Section 2.4.1: Preprocessing.

2.4.1. Preprocessing

The applied preprocessing is listed per the type of data and modelling phase:

- Vision System Data

The RGB data were preprocessed with autoscaling to uniformly model the variance among the different color channels.

- NIR spectra prior to PCA and MSPC

NIR spectra were pre-processed to remove effects, such as scattering, introducing variability not linked with information to be retrieved, and/or to enhance extractable information. In particular, Savitzky–Golay 2nd derivative and mean centering were applied prior to exploratory Principal Component Analysis and multivariate control charts building.

- NIR spectra prior to PLS regression

Savitzky–Golay 2nd derivative and mean centering were also used as preprocessing to compute the Partial Least Squares (PLS) regression model for the lipids content.

A different preprocessing strategy was needed to obtain the PLS model for consistency. This property was not directly linked to a chemical component, as the lipids show

specific absorption bands that can guide the modeling; thus, it was more difficult to model, especially considering how many registered on-line spectra were influenced by any process fluctuations. Thus, in order to remove spectral variability hindering the possibility of obtaining a satisfactory calibration model, a Dynamic Orthogonal Projection (DOP) [23] algorithm was applied, using the average spectra corresponding to the same consistency values (in the calibration set) as the source data (X_{source}) and the raw calibration spectra (X_{tar}) as the target. The main concept in DOP is that samples showing the same (or very close) y values should show the same spectral profile; thus, the "virtual" target spectra (X^*_{tar}), unaffected by the influence of uncontrolled conditions, could be estimated based on a distance or association matrix (M), calculated based on the y values of the source (y_s) and the target (y_t) domain. The singular value decomposition (SVD) of the difference matrix among measured and virtual target spectra was then used to determine the components (A) for orthogonalization:

$$X^*_{tar} = M^* \times X_{source} \quad (1)$$

$$D = X_{tar} - X^*_{tar} \quad (2)$$

$$[U_A\ S_A\ V_A] = svd(D, A) \quad (3)$$

$$X_{source\_corrected} = X_{source}\ (I - V_A V_A{}^T) \quad (4)$$

In our specific case, A = 4 was used after testing using from 1 to 5.

Once the average spectra were corrected, orthogonal projection could be directly used to predict the validation set, since the correction was embedded in the model. In this case, only mean centering (of both **X** and **y**) was applied prior to PLS.

2.4.2. Principal Component Analysis

Principal Component Analysis (PCA) is a method that by decomposition of the original data **X** into two matrices **T** and **P**, [24] according to Equation (5), allow reducing the dimensionality of the data set with a large set of variables, simplifying the exploration phase and the data visualization. PCA performs a projection of data from the original variables into new variables orthogonal to each other, the Principal Components (PCs), which are a linear combination of the original ones.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (5)$$

If the **X** matrix was composed of *n* rows (samples) and m columns (variables), the **T** matrix, called the scores matrix, which allowed us to understand the structure of the data, was composed by *n* rows and a number of columns equal to the number of PCs, and the loadings matrix **P** was composed by a number of rows equal to m and columns equal to the number of PCs. The loadings values corresponded to the weights by which each original variable entered the linear combination, thus defining the PCs, representing the contribution of each variable to each PC. The analysis of loadings matrix allowed us to understand the correlation structure of the variables [25]. The residual matrix **E**, which represented the unmodeled information, had the same dimension of **X,** and it was obtained by the subtraction of recalculated data from the PCA model ($\mathbf{TP}^T$) from **X**.

2.4.3. Multivariate Control Charts

PCA was also used to build multivariate control charts for MSPC. The dataset had been split in each calibration and test set manually, considering NOC observations, subdividing each period without production stops, as follows: the first part (about 65%) consisted of temporally contiguous points in the calibration set; and the second part (about 35%) was in the test set. In this way, we mimicked the real situation of continuous monitoring where samples to be predicted came after in time for each period. Observations not in NOC, as highlighted by exploratory PCA, were all included in the test set.

To estimate the correct number of PCs, cross-validation was performed with a *venetian blind* scheme with ten splits. The MSPC charts were based on two parameters: Hotelling T^2, which described the distance of a sample in the model space, and Q, which defined the distance of a sample from the model space. In other words, if a sample had high T^2 values, the model was able to describe it, but the distance between the sample and the center of the model was high, i.e., it showed an extreme behavior. On the other hand, if a sample was characterized by high Q values, the model was not able to describe the sample properly, hence the correlation structure of variables was different from the other samples. To assess if a sample was extreme or anomalous, signifying a departure from normal operative conditions for both control charts, the acceptance limits had to be estimated. The T^2 limit was obtained based on Hotelling's T^2 distribution, whereas the Q limit was based on χ^2 distribution and was calculated either with Jackson and Mudholkar approximation or the Box method [26,27].

2.4.4. PLS Regression

PLS is a linear regression method that allows predicting one or more response variables (Y block) from a predictor matrix (X block), establishing a multivariate linear relationship. It operates in a low-dimensional space defined by the Latent Variables (LVs), obtained from the simultaneous decomposition of **X** and **Y**, which are oriented on directions of maximum covariance between **X** and **Y** [28]. A PCA-like decomposition of **X** and **Y** is achieved (outer relation):

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{6}$$

$$\mathbf{Y} = \mathbf{U}\,\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{7}$$

where an inner relation links the outer relation:

$$\mathbf{U} = b^{*}\mathbf{T} \tag{8}$$

Hence, re-expressing this as a regression model:

$$\grave{\mathbf{Y}} = \mathbf{X}\,\mathbf{B} \tag{9}$$

where **T** and **U** are **X** and **Y** scores, **P** and **Q** are **X** and **Y** loadings, and **E** and **F** are the residual matrices, respectively. B holds the regression coefficients that allow the prediction of **Y** from **X** directly.

Data were partitioned into calibration (70%) and validation (30%) sets by the means of a Duplex algorithm [29]. The PLS model dimensionality, i.e., the number of PLS components, was assessed by the Root Mean Square Error in Cross-Validation (RMSECV), while the Root Mean Square Error in Prediction (RMSEP) was used to evaluate the models' predictive capability. Residual plots were also inspected.

3. Results and Discussion

3.1. Exploratory Data Analysis

Each type of data, RGB parameters, and NIR spectra were analyzed separately to visualize and explore the data structure. PCA analysis carried out on NIR spectra (acquired for 459 time points) had highlighted the presence of a cluster of samples at the negative value of PC1 and positive value of PC2, as shown in Figure 2a, as very far and different from all the other samples. Observing the PC1 versus time plot (Figure 2b), it was evident that these samples always corresponded to restarts, where production started after a period of inactivity. In Figure 2c, the loadings line plots for PC1 and PC2 are shown as the blue and red lines, respectively, where it is possible to see the absorption bands as mainly responsible for this difference. However, to jointly interpret scores and loadings plots, a PC1 vs. PC2 loadings scatter plot was also generated (Figure 2d). In the two figures, highlighted in purple, the wavelengths that describe the separation between the NOC and anomalous

samples are shown. It can be observed that the band in PC1 at 1400 nm, despite being the most intense, is not involved in the description of anomalous samples but just in extreme NOC samples with high values of PC1 scores in Figure 2a. On the other hand, the bands at 1170, 1213, 1236, and 1410 nm describe the behavior of the anomalous samples, as they fell in the separation direction, meaning that these samples had very different absorptions at these wavelengths. In detail, the bands at 1178 and 1410 nm can be ascribable to lignin, namely, the second overtone of C-H bond stretching of CH_3, and to the first overtone of the O-H bond stretching of the ROH group, respectively. Whereas, the band at 1213 and 1236 nm are related to the first and second overtone of C-H bond stretching of oleic and linoleic acid in olive oil CH_2 [30,31].

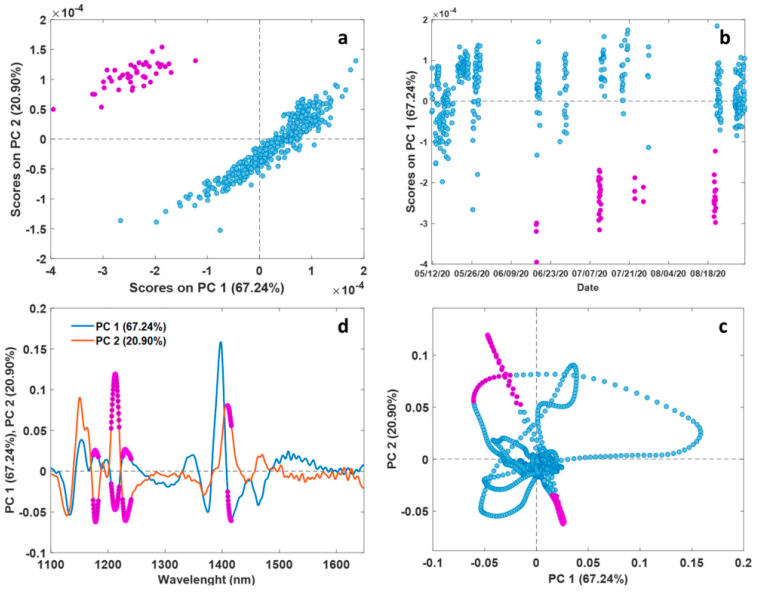


Figure 2. Results of the Exploratory Data Analysis performed on NIR data. PC1 vs. PC2 Scores plot (**a**), Scores on PC1 as a function of time (**b**), Loadings on PC1 and PC2 as a function of time (**c**), and Loadings on PC1 vs. PC2 (**d**). In (**a**,**b**), purple points represent anomalous samples; in (**c**,**d**), purple points represent wavelengths that mainly depict the difference of anomalous samples from the other ones.

Since these samples show the outliers' behavior, as they clearly do not represent the Normal Operative Conditions (NOCs), they were removed, and a new PCA model was built in order to obtain a better visualization of the possible differences among NOC samples.

The first PC (79.36% of variance explained) did not show any interesting trend, thus PC2 and PC3 were inspected. In Figure 3a,b, the scores plot of PC2 vs. PC3 is reported, where samples are colored according to the different additional information available, i.e., suppliers and different cuts, respectively. The suppliers' names have not been disclosed because of confidential agreement restrictions with the company. PC2 discriminated samples according to suppliers, as almost all samples of supplier number two had positive PC2 values, and the samples of suppliers three and four had negative PC values, suggesting that they were more similar to each other, with respect to number two. Only the samples coming from supplier five did not clearly differentiate from the others, whereas the number of samples from supplier one were too low to judge. Furthermore, PC2 and PC3 could distinguish between samples related to cut one and two (negative values of PC2 and positive values of PC3), with respect to samples related to cut three and four. The possibility to

discriminate against different cuts is relevant for the company, as younger basil plants generally give a higher quality product. However, observing the two plots simultaneously, it is evident that only certain suppliers, namely, number three and four, had delivered samples characterized by low cuts. In Figure S1a,b, the loadings plots of PC2 and PC3 are reported, respectively, which show the NIR bands responsible for these differences. Even if it is not possible to assess if suppliers or cuts influence them, the PCA resulted in a valuable tool to assess if incoming information about raw materials could be linked to the intermediate product characteristics; evidently, a more systematic planning of the next harvesting campaigns could clarify if a cut or supplier were the influential factors.

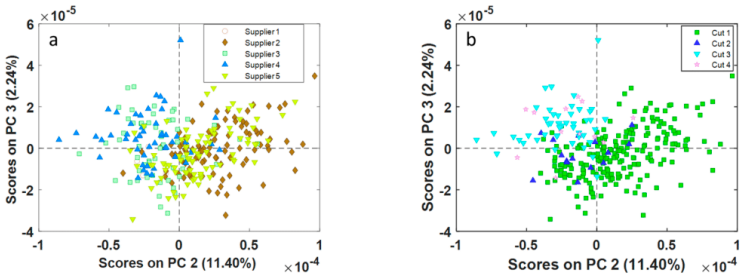


Figure 3. Results of the Exploratory Data Analysis performed on NIR data. PC2 vs. PC3 scores plots colored by different suppliers (**a**) and cuts (**b**).

PCA analysis carried out on data collected by an RGB camera was not able to detect the anomalous behavior of the samples highlighted in Figure 1. A possible explanation is that the process needed time to return to NOCs after a stop when the production restarted, and it could happen that the NIR spectra referred to material that was probably a residue of the old process (before the restart), and thus the acquired spectra did not depict the intermediate product newly produced at the beginning. Moreover, the observation of the samples' separation due to different cuts or suppliers was less efficient than the respective analysis performed on the NIR spectra. Thus, these differences were not linked to color variation but mostly to the basil's "chemical" profile.

3.2. MSPC Charts

The most interesting results related to the MSPC charts based on PCA were obtained by using the NIR data only (inclusion of RGB parameters did not provide additional insights). The PCA model, which explains 93% of the data variance with 4 Principal components, was calculated by inserting only the samples that were considered in NOCs according to plant experts in the calibration set (294 samples), whereas the test set (165 samples) comprised both NOCs and anomalous samples. The T^2 chart, reported in Figure 4a, describes the distance of each sample from the origin within the model space. Black circles represent the calibration samples used to build the PCA model, whereas red diamonds represent the test samples projected on the model. This chart detected five groups of samples with high T^2 values, which, again, corresponded to the NIR spectra acquired at the different restarts of the production. No other test sample exceeded the T^2 limit. Regarding the Q chart (Figure 4b), which describes the distance of each sample from the model space, the same samples corresponding to the restart are seen anomalous as for the T^2 chart, meaning that the model did not properly describe these samples. The charts' limits include few non-consecutive samples and inside of the nominal 5% of the total.

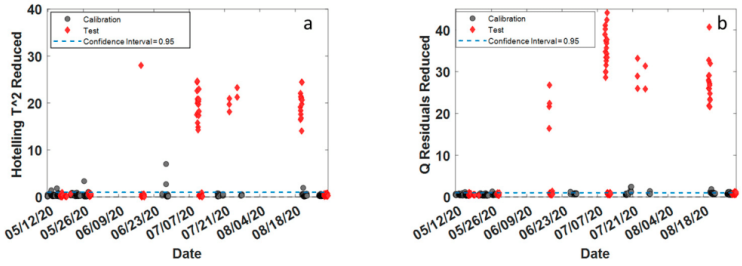


Figure 4. T^2-(**a**) and Q-(**b**) based MSPC charts.

Samples were also colored according to cut, supplier, consistency, and lipids values to observe if their behavior was related to these different features, but no particular trends were detected.

Nonetheless, the results obtained show how these charts are efficient in detecting possible departure from NOC, which translate to differences in intermediate products, accelerating the identification of possible plant issues or, as in this case, the adaptation of the process while returning to NOCs after a stop period. NIR is a very sensible technique to signal any variability occurring in intermediate production samples that can be due to process resetting (actual case), process drift, or variation in the NIR instrumentation setting/performance. The interpretation of the loadings and analysis of previous production campaigns data may help in discerning the different situations.

3.3. Predictive Models

An attempt to obtain predictive models, which can then be possibly used to estimate the consistency and lipids content of the final product in real time, was undertaken. Since RGB data were not able to provide reliable prediction models for both parameters, only results obtained by NIR data are presented, as summarized in Table 1.

Table 1. Results obtained by PLS Regression.

Method	LVs	RMSECV	RMSEP
Consistency (cm)	9	0.64	0.68
Lipids (%)	5	1.59	2

Before model computation, data were split by using a duplex algorithm with a 70/30% proportion in the calibration and test sets, giving 142(cal)/61(test) and 33(cal)/12(test) for consistency and lipids, respectively. Afterwards, four samples belonging to the anomalous group of observations, detected by using the T^2 and Q distances, were removed from the test set for consistency.

The prediction model for consistency was built using 9 LVs, corresponding to the minimum RMSECV (*venetian blind*, 10 splits) value. The RMSEP value was close to RMSECV (Table 1) and corresponded to an average relative percentage error of 10% in prediction, which was considered acceptable by the company for an early (intermediate product) on-line quality estimation. The samples in the test set showed a rather high variability compared to the ones in the calibration (Figure 5a,b). Nonetheless, the residuals vs. measured values of the consistency plot (Figure 5b) highlighted that the errors on both the calibration and test samples were randomly distributed, not showing any visible trend, excluding any bias.

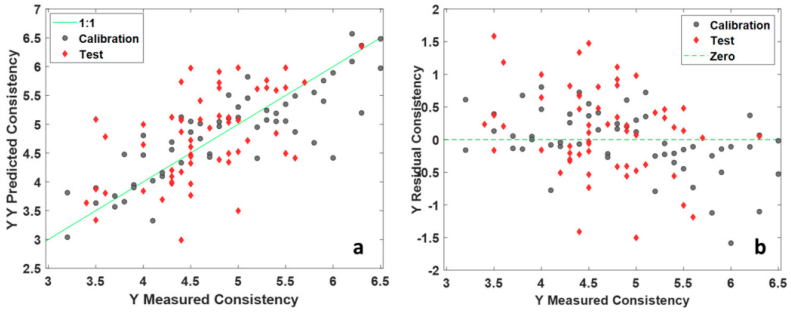


Figure 5. PLS results on NIR data for consistency. Predicted vs. measured values plot (**a**), residuals vs. measured values plot (**b**).

The prediction model of total lipids content was built using a lower number of samples than the previous model, as this parameter was assessed less frequently than consistency. In this case, 5 LVs were selected according to the minimum RMSECV (*venetian blind*, 10 splits) for the model's construction. As shown in Figure 6a, the majority of the samples had a lipid content included in the range 46–49%, and only a few samples presented higher values. This is a quite common situation in real time production, where a consistent quality of the product is pursued. In this case, a couple of samples in the test set were predicted with a higher error but, in general, the error values comprised the 2% range, which the company considered acceptable for controlling if the product was within specification for this parameter. One of the two samples with a high lipid content in the test set was predicted accurately, whereas the other one was underestimated (Figure 6b).

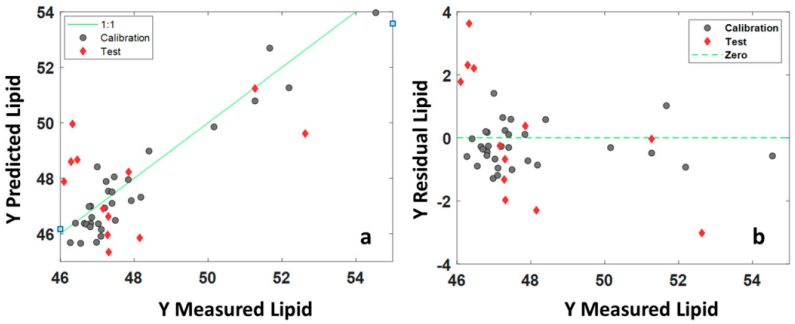


Figure 6. PLS results on NIR data for lipids content. Predicted vs. measured values plot (**a**), residuals vs. measured values plot (**b**).

In Figure 7, the Variable Influence in Projection (VIP) scores are shown [32], which highlight that the band at 1166 nm, ascribable to the olive oil's second overtone of the CH stretching of CH_3 [30,31], is the most influent for the prediction of total lipids content. Moreover, other bands linked to lipids in olive oil [30,31] can be found at 1422 and 1461 nm, typical of the CH stretching and deformation of CH_2, both above the significance threshold [32].

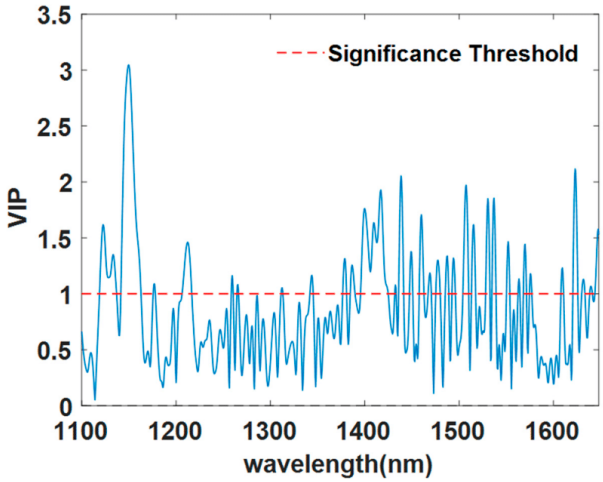


Figure 7. VIP scores of PLS on NIR data for lipids content.

4. Conclusions

This study presents a feasibility study towards the real-time monitoring of an industrial food process line (pesto production). Since historical data were not available, the obtained results referred to a single basil harvesting campaign. The modeling effort concerned both latent variables based multivariate control charts, aimed at monitoring the stability of process conditions and the eventual detecting of fluctuations exceeding the natural variability of the process as well as the quality properties' prediction in real time. Despite the fact that the collected data were limited, the results gave interesting insights, which are summarized in the following.

4.1. MSPC Results

(i) the RGB parameters obtained by the vision system, albeit potentially very useful, were not increasing information retrieved from NIR. We think this is due to the limited number of features extracted by the image, which could otherwise provide a good characterization of raw material; further work is in progress in this direction (e.g., detecting the percentage areas of damaged leaves, branches, and stems by an image analysis tool);

(ii) NIR-based multivariate control charts could detect restarts after temporary production stoppages, underlining that some changes occur in the intermediate product. On one hand, this is an indication of how sensible NIR spectroscopy is to monitor any changes, and, on the other hand, a monitoring system can clearly indicate when process fluctuations return to natural process variabilities and to the constancy of the product.

4.2. On-Line Predictive Models

(iii) The predictive model to estimate the pesto's consistency and total lipids content, based on the NIR spectra of the intermediate product, gave errors in the external predictions, which are considered acceptable by the company for on-line quality estimation.

(iv) It is worth noting that while building predictive models of final product quality parameters based on on-line sensors data is highly desirable, they suffer from the limited response variability (which, evidently, should be confined in the in-specific ranges). When, as in this case, it is not possible to expand the calibration range by pilot studies, the models can be, nonetheless, used as a timely rough indication of the property's value. In this respect, more than an estimation of the quality values, they may give a preliminary check about respecting specifications Within this framework, the obtained models seem promising.

Finally, it is worth mentioning the main issues encountered, such as the lack of systematic recording of acquired on-line data, the difficulties in recovering a sound synchronization scheme, and the critical role of spectral preprocessing to cope with the many sources of variabilities intrinsic in a process framework.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/foods12081679/s1, Figure S1: Results of the Exploratory Data Analysis performed on NIR data. Loadings Plot (a) PC1 vs. wavelengths; (b) PC2 vs. wavelengths.

Author Contributions: Conceptualization, D.T., A.D., M.C. and L.S.; methodology, D.T., C.D., M.C. and L.S.; software, D.T. and L.S.; validation, D.T., A.D., C.D. and L.S.; formal analysis, M.C. and L.S.; investigation, D.T., S.T. and A.D.; resources, A.D. and M.C.; data curation, D.T., S.T. and A.D.; writing—original draft preparation, D.T., C.D. and L.S.; writing—review and editing, D.T., A.D. M.C. and L.S.; visualization, D.T. and S.T.; supervision, C.D., M.C. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are unavailable due to privacy restrictions.

Acknowledgments: L. Strani aknowledges R. Vitale (University of Lille) for suggestions and support in the course of a Virtual Mobility Grant in the frame of the COST Action CA19145 "European Network for Assuring Food Integrity using Non-Destructive Spectral Sensors (SENSORFINT).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ferrer-Riquelme, A. Statistical Control of Measures and Processes. In *Comprehensive Chemometrics*; Elsevier: Amsterdam, The Netherlands, 2009; pp. 97–126.
2. Grassi, S.; Strani, L.; Alamprese, C.; Pricca, N.; Casiraghi, E.; Cabassi, G. A FT-NIR process analytical technology approach for milk renneting control. *Foods* **2022**, *11*, 33. [CrossRef] [PubMed]
3. Franca, L.; Grassi, S.; Pimentel, M.F.; Amigo, J.M. A single model to monitor multistep craft beer manufacturing using near infrared spectroscopy and chemometrics. *Food Bioprod. Process.* **2021**, *126*, 95–103. [CrossRef]
4. Zhou, Q.; Dai, Z.; Song, F.; Li, Z.; Song, C.; Ling, C. Monitoring black tea fermentation quality by intelligent sensors: Comparison of image, e-nose and data fusion. *Food Biosci.* **2023**, *52*, 102454. [CrossRef]
5. Catelani, T.A.; Santos, J.R.; Páscoa, R.N.; Pezza, L.; Pezza, H.R.; Lopes, J.A. Real-time monitoring of a coffee roasting process with near infrared spectroscopy using multivariate statistical analysis: A feasibility study. *Talanta* **2018**, *179*, 292–299. [CrossRef]
6. Hao, Y.; Lu, Y.; Li, X. Study on robust model construction method of multi-batch fruit online sorting by near-infrared spectroscopy. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2022**, *280*, 121478. [CrossRef]
7. Strani, L.; Grassi, S.; Alamprese, C.; Casiraghi, E.; Ghiglietti, R.; Locci, F.; Pricca, N.; De Juan, A. Effect of physicochemical factors and use of milk powder on milk rennet-coagulation: Process understanding by near infrared spectroscopy and chemometrics. *Food Control* **2021**, *119*, 1074. [CrossRef]
8. Maléchaux, A.; Le Dréau, Y.; Artaud, J.; Dupuy, N. Control chart and data fusion for varietal origin discrimination: Application to olive oil. *Talanta* **2020**, *217*, 121115. [CrossRef]
9. Pérez-Beltrán, C.H.; Jiménez-Carvelo, A.M.; Torrente-López, A.; Navas, N.A.; Cuadros-Rodríguez, L. QbD/PAT—State of the Art of Multivariate Methodologies in Food and Food-Related Biotech Industries. *Food Eng. Rev.* **2023**, *15*, 24–40. [CrossRef]
10. Grassi, S.; Alamprese, C. Advances in NIR spectroscopy applied to process analytical technology in food industries. *Curr. Opin. Food Sci.* **2018**, *22*, 17–21. [CrossRef]
11. Pu, Y.Y.; O'Donnell, C.; Tobin, J.T.; O'Shea, N. Review of near-infrared spectroscopy as a process analytical technology for real-time product monitoring in dairy processing. *Int. Dairy J.* **2020**, *103*, 104623. [CrossRef]
12. Baines, T.; Brown, S.; Benedettini, O.; Ball, P.D. Examining green production and its role within the competitive strategy of manufacturers. *J. Ind. Eng. Manag.* **2012**, *5*, 53–87. [CrossRef]
13. Rico-Rodriguez, F.; Strani, L.; Grassi, S.; Lancheros, R.; Serrato, J.C.; Casiraghi, E. Study of Galactooligosaccharides production from dairy waste by FTIR and chemometrics as Process Analytical Technology. *Food Bioprod. Process.* **2021**, *126*, 113–120. [CrossRef]
14. Strani, L.; Mantovani, E.; Bonacini, F.; Marini, F.; Cocchi, M. Fusing NIR and Process Sensors Data for Polymer Production Monitoring. *Front. Chem.* **2021**, *9*, 785. [CrossRef] [PubMed]
15. Westerhuis, J.A.; Gurden, S.P.; Smilde, A.K. Generalized contribution plots in multivariate statistical process monitoring. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 95–114. [CrossRef]

16. Avila, C.; Mantzaridis, C.; Ferré, J.; de Oliveira, R.R.; Kantojärvi, U.; Rissanen, A.; Krassa, P.; De Juan, A.; Muller, F.L.; Hunter, T.; et al. Acid number, viscosity and end-point detection in a multiphase high temperature polymerisation process using an online miniaturised MEMS Fabry-Pérot interferometer. *Talanta* **2021**, *224*, 121735. [CrossRef] [PubMed]
17. Macho, S.; Rius, A.; Callao, M.P.; Larrechi, M.S. Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy: Standardisation of the calibration model. *Anal. Chim. Acta* **2001**, *445*, 213–220. [CrossRef]
18. Joshi, K.; Patil, B. Multivariate statistical process monitoring and control of machining process using principal component-based Hotelling T2 charts: A machine vision approach. *Int. J. Product. Qual. Manag.* **2022**, *35*, 40–56. [CrossRef]
19. Biancolillo, A.; Scappaticci, C.; Foschi, M.; Rossini, C.; Marini, F. Coupling of NIR Spectroscopy and Chemometrics for the Quantification of Dexamethasone in Pharmaceutical Formulations. *Pharmaceuticals* **2023**, *16*, 309. [CrossRef]
20. de Oliveira, R.R.; Pedroza, R.H.; Sousa, A.O.; Lima, K.M.; de Juan, A. Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy. *Anal. Chim. Acta* **2017**, *985*, 41–53. [CrossRef]
21. Kourti, T. Application of latent variable methods to process control and multivariate statistical process control in industry. *Int. J. Adapt. Control Signal Process.* **2005**, *19*, 213–246. [CrossRef]
22. Strani, L.; Vitale, R.; Tanzilli, D.; Bonacini, F.; Perolo, A.; Mantovani, E.; Ferrando, A.; Cocchi, M. A Multiblock Approach to Fuse Process and Near-Infrared Sensors for On-Line Prediction of Polymer Properties. *Sensors* **2022**, *22*, 1436. [CrossRef]
23. Zeaiter, M.; Roger, J.M.; Bellon-Maurel, V. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 227–235. [CrossRef]
24. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
25. Bro, R.; Smilde, A.K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [CrossRef]
26. Jackson, J.E.; Hearne, F.T. Hotelling's T_M^2 for Principal Components—What about Absolute Values? *Technometrics* **1979**, *21*, 253–255.
27. Nomikos, P.; MacGregor, J.F. Multivariate SPC charts for monitoring batch processes. *Technometrics* **1995**, *37*, 41–59. [CrossRef]
28. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
29. Snee, R.D. Validation of regression models: Methods and examples. *Technometrics* **1977**, *19*, 415–428. [CrossRef]
30. Galtier, O.; Dupuy, N.; Le Dréau, Y.; Ollivier, D.; Pinatel, C.; Kister, J.; Artaud, J. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. *Anal. Chim. Acta* **2007**, *595*, 136–144. [CrossRef]
31. Casale, M.; Simonetti, R. Near infrared spectroscopy for analysing olive oils. *J. Near Infrared Spectrosc.* **2014**, *22*, 59–80. [CrossRef]
32. Wold, S.; Johansson, E.; Cocchi, M. PLS: Partial least squares projections to latent structures. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kluwer ESCOM Science Publisher: Dordrecht/Leiden, The Netherlands, 1993; pp. 523–550.

chemosensors

MDPI

Article

Near Infrared and UV-Visible Spectroscopy Coupled with Chemometrics for the Characterization of Flours from Different Starch Origins

Samuele Pellacani, Marco Borsari, Marina Cocchi, Alessandro D'Alessandro, Caterina Durante *, Giulia Farioli and Lorenzo Strani

Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy; samuele.pellacani@unimore.it (S.P.); marco.borsari@unimore.it (M.B.); marina.cocchi@unimore.it (M.C.); alessandro.dalessandro@unimore.it (A.D.); 255751@studenti.unimore.it (G.F.); lorenzo.strani@unimore.it (L.S.)

* Correspondence: caterina.durante@unimore.it

Abstract: This work tested near-infrared (NIR) and UV-visible (UV-Vis) spectroscopy coupled with chemometrics to characterize flours from different starch origins. In particular, eighteen starch-containing flours (e.g., type 00 flour, rye, barley, soybean, chestnut, potato, spelt, buckwheat, oat, millet, rice, durum wheat, amaranth, chickpea, sesame, corn, hemp and sunflower flours) were analyzed with a twofold objective: chemically characterizing the investigated flours and laying the groundwork for the development of a fast and suitable method that can identify the botanical source of starch in food. This could ensure ingredient traceability and aid in preventing/detecting food fraud. Untargeted approaches were used for this study, involving the simultaneous acquisition of a large amount of chemical information (UV-Vis on extracted starch and NIR signals on raw flours) coupled with chemometric techniques. UV-VIS spectra were acquired between 225 and 800 nm after sample pretreatment to extract starch. NIR spectra were acquired between 900 and 1700 nm using a poliSPEC NIRe portable instrument on the flours without any kind of pretreatments. An initial exploratory investigation was conducted using principal component analysis and cluster analysis, obtaining interesting preliminary information on patterns among the investigated flours. In particular, the UV-Vis model successfully discerned samples such as potato, chestnut, sunflower, durum wheat, sesame, buckwheat, rice, corn, spelt and 00-type flours. PCA model results obtained from the analysis of NIR spectra also provided comparable results with the UV-Vis model, particularly highlighting the differences observed between hemp and potato flours with soybean flour. Some similarities were identified between other flours, such as barley and millet, rye and oats, and chickpea and amaranth. Therefore, some flour samples underwent surface analysis via scanning electron microscope (SEM) using the Nova NanoSEM 450 to detect distinctive morphology.

Keywords: starch; UV-Vis spectroscopy; NIR spectroscopy; principal component analysis; cluster analysis

Citation: Pellacani, S.; Borsari, M.; Cocchi, M.; D'Alessandro, A.; Durante, C.; Farioli, G.; Strani, L. Near Infrared and UV-Visible Spectroscopy Coupled with Chemometrics for the Characterization of Flours from Different Starch Origins. *Chemosensors* **2024**, *12*, 1. https://doi.org/10.3390/chemosensors12010001

Academic Editors: Xiong Wan, Lei Zhang and Jiulin Shi

Received: 17 November 2023
Revised: 13 December 2023
Accepted: 20 December 2023
Published: 22 December 2023

1. Introduction

Starch is one of the major natural polysaccharides and is widely used in many areas of industry [1–6]. Annually, approximately 60 million tons of starch are extracted worldwide from various cereal, tuber, and root crops. It finds widespread use across industries, with 60 percent utilized in the food sector (including baked goods, sauces, soups, and confectionery) and the remaining 40 percent in pharmaceuticals and non-edible products like fertilizers, paper, cardboard, and packaging [7].

The botanical origin of the starch used as a raw material is a pivotal determinant of the processing and overall quality of the final product [7–10], confirming the need for a rapid, inexpensive, and reliable method for its determination [10–12].

Chemically, starch consists of two glucose polymers, amylose and amylopectin, both of which differ in structure and size. Amylose is a predominantly linear polymer with α-(1,4) glycosidic bonds and has a relatively small structure (up to 106 Da). On the other hand, amylopectin is a highly branched polymer formed by linear chains with different degrees of polymerization and has a much higher degree of α-(1,6) glycosidic bonds [10,12,13]. The botanical origin determines the ratio and association of amylose and amylopectin as well as their morphological structure as granules with variable size (1–100 μm) and shape. Amylose and amylopectin molecules exhibit distinct structures, which are influenced by the botanical source of the starch. Factors such as the molecular size, inner chain length, and presence of side chains contribute to these characteristic differences. For instance, amylose derived from wheat is composed of a limited quantity of expansive, branched molecules. In contrast, sweet potato amylose consists of a small number of relatively large, unbranched molecules [7,8,11].

There are several methodologies in the literature for starch analysis and its quantification [10–16]. Most studies involve pretreatment of the sample by gelatinizing starch at elevated temperatures in the presence of a thermostable α-amylase enzyme to produce a series of linear and branched dextrins, which are subsequently hydrolyzed into glucose. In addition, there are two alternatives to the enzymatic conversion of starch to glucose: (i) the dissolution of starch with hydrochloric acid and (ii) the dissolution of starch with a boiling solution of calcium chloride.

The analytical methodologies commonly employed for ascertaining the botanical origin of starch typically encompass indirect techniques, including optical and electron microscopy, as well as analytical methods reliant on (i) enzymatic reactions, (ii) chromatography, (iii) X-ray diffraction, and various spectroscopic techniques [10–18]. Nonetheless, these approaches are time-consuming and necessitate specific pretreatments of the sample, such as starch extraction and subsequent analysis.

In this context, UV-Vis spectroscopy can also be used due to the ability of amylose and amylopectin to form blue-colored helical inclusion complexes with the triiodide ion. Consequently, UV-Vis spectroscopy leverages the unique absorption spectrum of amylose and amylopectin–triiodide complexes, acting as a sample fingerprint [11].

Among the fast and non-invasive analytical techniques, near-infrared (NIR) spectroscopy can be particularly suitable for determining the chemical composition of components in a variety of complex organic samples [18–21]. Recent advancements in instrumentation, miniaturization, wireless communication systems, and sophisticated algorithms dedicated to statistical data processing have facilitated the development of numerous applications in various research fields, enabling an at-line, on-line, and in-line non-destructive analysis of a wide array of food products.

NIR spectra can exhibit several bands, including those attributed to the O-H stretch first overtone, which is characterized by absorbance peaks at 1450 nm and 1540 nm, respectively [19,21]. Owing to the presence of these distinctive bands, NIR spectroscopy has been successfully employed as a rapid and non-destructive technique for quantifying the starch content in various types of flours [19].

The aim of the present study was to assess the feasibility of NIR and UV-Vis spectroscopy in combination with chemometrics for distinguishing different botanical origins of starch. These methods proved to be particularly suitable for implementation in laboratories equipped with basic UV-Vis or NIR spectrophotometers. Additionally, the NIR-based method can be directly applied in situ by portable NIR devices, such as the one used in this study. Utilizing a portable NIR instrument at flour delivery would enable the analysis of all batches entering production in a more representative and efficient manner.

Among the flours chosen, corn starch accounts for 80% of the worldwide market [1]. Barley, corn, potato and rice flour are considered conventional sources of starch. On the other hand, amaranth, buckwheat, chestnut, chickpea and millet are considered non-conventional and emerging sources of starch. In particular, chestnut starch presents a pasting profile similar to corn one, making it a potential alternative to corn starch [1].

In pursuit of the objectives of this research, an untargeted approach was adopted, utilizing the entire acquired signals from the different used analytical techniques, namely spectroscopic techniques, as a fingerprint of the investigated samples. The advantage of this approach lies in its ability to provide a broad spectrum of information for data processing and mining without a predefined set of compounds. The obtained spectroscopic signals served as a comprehensive fingerprint of the investigated samples and coupled with chemometrics analysis could allow the identification of patterns among the investigated flours based on their starch origin. For UV-Vis spectroscopy, the complete recorded spectrum (190 to 800 nm) was considered, encompassing not only the signals associated with the triiodide–starch complex (bands between 400 and 800 nm). In the case of NIR, no pre-treatment was applied, resulting in a spectrum influenced not solely by starch but also by water, lipids, and proteins.

An exploratory analysis was performed on all the obtained signals using principal component analysis (PCA) [22]. PCA was employed to process all acquired signals and provide insights into the presence of similarities and differences among the examined samples. Furthermore, UV-Vis and a selected region of NIR spectra were separately used as a unique fingerprint and elaborated by cluster analysis [23] in order to verify if it is possible to distinguish starches of different botanical origin.

Additionally, a subset of flour samples, identified as similar by chemometric analysis, underwent surface analysis through scanning electron microscopy (SEM) using the Nova NanoSEM 450 microscope to identify potential distinctive morphological features. This choice was made to support the proposed methods when the objective is to be more selective in the identification of similar starches with different botanical origin.

The physical and chemical properties of starch have been discussed in detail in the literature [13], but as far as the botanical origin of starch is concerned, only a few studies are present where were explored the use of spectrophotometric techniques, specifically UV-Vis. The present approach allows for a more thorough and complete characterization of flours, marking the first case where both UV-Vis and NIR spectroscopy have been employed on a diverse and large range of flour types. The incorporation of chemometric techniques improves the interpretability and depth of the analysis, allowing a thorough understanding of spectral information and its implications for flour characterization. Furthermore, the study integrates the developed models with data related to the morphology of certain flours, acquiring SEM images on raw flours that exhibited similarities in the chemometric analysis. This multidimensional integration of spectroscopic data and morphological information offers a more holistic and comprehensive perspective on flour characterization.

The ability to determine the botanical origin of starch is crucial for the food industry, especially in the quality verification of raw materials for baked goods. It ensures ingredient traceability and aids in preventing food adulteration and fraud.

2. Materials and Methods

2.1. Samples and Reagents

Eighteen flours samples, type 00 flour, rye, barley, soybean, chestnut, potato, spelt, buckwheat, oat, millet, rice, durum wheat, amaranth, chickpea, sesame, corn, hemp and sunflower flours, were purchased from the market and were stored inside sealed polyethylene containers to preserve them from possible contamination.

Potassium triiodide solution was prepared using iodine and potassium iodide, which were both purchased from Sigma–Aldrich, Merck, Darmstadt, Germany.

2.2. Starch Sample Preparation for UV Analysis

In order to acquire the starch fingerprint spectrum by the UV-Vis technique, a preliminary treatment of the flour sample was necessary to extract the starch. Based on the literature [11], the following analytical procedure was developed and subsequently applied to all the investigated samples. A sample aliquot of 1.25 g of flour was dispersed in 50 mL of deionized water and heated to boiling for 15 min. Then, 10 mL of supernatant was taken

with a sterilized syringe and transferred inside a 25 mL volumetric flask. Afterwards, 2 mL of a potassium triiodide solution 0.03 M was added under stirring, and the volumetric flask was filled to the mark with deionized water.

The blank sample, used as a reference for spectroscopic measurements, was obtained by diluting 2 mL of potassium triiodide solution inside a 25 mL volumetric flask with deionized water.

For each starch type, three independent solutions were prepared obtaining three UV-Vis spectra for each sample. Chickpea flour was selected as a control sample to monitor the reproducibility of the analytical method and was analyzed five times. The decision to include a control sample was driven by the intention to closely monitor the performance of the portable NIR device over the different experimental sessions. It is of utmost of importance to select a control sample that shares similar characteristics with the diverse range of samples under investigation and possesses the ability to maintain consistent chemical and physical properties over time. This ensures that any observed differences among control samples can be attributed solely to variations in measurements and not to inherent sample instability. In this case, the choice of any sample as a control would have been suitable given that all the samples in the present study exhibited stable chemical and physical properties. Therefore, the selection of chickpea flour as the control was influenced by the practical consideration of sample availability, since its quantity at our disposal exceeded that of other potential control samples.

2.3. Spectrophotometric Measurements

UV-Vis spectra were recorded with a JASCO V-750 UV/Vis/NIR spectrophotometer (JASCO, Tokyo, Japan) at 298 K in a 225–800 nm spectral range employing quartz cells (1 cm optical path) with a resolution of 1 nm.

2.4. Near-Infrared (NIR) Spectroscopy

Approximately 100 mg of flour was inserted into a sample holder (plastic vessel) with almost the same thickness (less than 4 mm) to avoid any scattering in the acquisition. Next, three NIR spectra were acquired for each vessel by placing the instrument at three different points: at the top, middle and bottom of the vessel.

While powder samples are generally assumed to be homogeneous, the decision to conduct measurements at multiple points within the vessel stems from the common practice of acquiring replicate measurements. This approach is employed not only to enhance the representativeness of the measurements but also to address specific challenges associated with the use of portable or miniaturized NIR instruments [24]. In the context of our research, a portable NIR instrument has also been utilized, as it offers the advantage of developing an analytical method that can be directly applied in situ. This is particularly valuable, for example, in a commercial setting where monitoring different batches of flour deliveries is essential. It is well documented in the literature that while miniaturized instruments yield satisfactory results, they may exhibit lower performance compared to benchtop instruments [24]. Factors such as lower representative transmission spectra and spectral resolution contribute to these differences [24]. Furthermore, when dealing with powder samples, a container is often necessary for measurements. It is crucial to ensure that the chosen container does not introduce any artifacts or affect the sample spectra. Given these considerations and the use of plastic vessels in conjunction with a portable NIR instrument in our study, acquiring replicates at different points within the samples inside the holder was deemed necessary to ensure the representativeness of the sampling.

In order to investigate the reproducibility of the analytical method and check for any systematic errors, a sample of flour (chickpea flour) was selected as a control sample, and it was analyzed three times per measurement session (at the beginning, at the middle and at the end of the session) following the procedure previously described, obtaining 27 replicates.

The NIR spectra were collected using a portable NIR spectrometer, poliSPEC NIRe (ITPhotonics S.r.l., Fara Vicentino, Italy). Due to its diffraction grating and the double chip InGaAs 512 pixels sensor with a controlled cooling system, poliSPEC NIRe covers the spectral range of 930–1700 nm with an average numerical resolution of 3.2 nm and an average optical resolution HWHM of 3.25 nm.

2.5. Scanning Electron Microscopy

Small amounts of sample were sufficient to acquire images, which were attached to a circular holder (stub) using double-sided adhesive tape. Prior to SEM investigation, a metallization process was carried out, during which the sample surface was coated with a thin layer of gold (20 nm/min for 45 s) to improve the electrical conductivity of the sample and to preserve the sample morphology. Observation of the samples was performed using the Nova NanoSEM 450 (Fei Company-Bruker corporation) scanning electron microscope. The scanning electron microscope was operated at an accelerating voltage (high voltage, HV) of 20 kV, and images were obtained using the directional backscatter detector.

2.6. Data Analysis

UV-Vis and NIR spectral data were imported and processed under MATLAB 2020a (The MathWorks, Inc., Natick, MA, USA) environment. Signal preprocessing, PCA and cluster analysis were performed by PLS-Toolbox v. 8.9 (Eigenvector Inc., Manson, WA, USA).

The assignment of NIR flour signals involved a comparative analysis with existing literature [19,20,25]. Prior to the development of chemometric models, UV-Vis spectra underwent mean centering, while NIR spectral data were pre-processed using a standard normal variate (SNV) [26] to mitigate the baseline shift, noise, and the impact of light scatter. Initially, principal component analysis (PCA) [22] was conducted on the pre-processed spectra to explore the data and identify potential similarities and differences among flour samples.

Subsequently, cluster analysis [23] was employed to attempt the identification of groups without relying on pre-established class memberships. Most cluster analysis methods assume that samples close together in the measurement space are similar and likely belong to the same class. Various ways exist to define the distance between samples with the Euclidean distance being the most common. It is calculated as the square root of the sum of squared differences between the samples. In this paper, considering the multivariate nature of the data, the Euclidean distance was computed by taking into account the scores on all the principal components (PCs) of the model. Specifically, the distance (d_{ij}) between samples x_i and x_j with scores t_i and t_j was defined as follows:

$$d_{ij} = \sqrt{(t_i - t_j)(t_i - t_j)^T} \quad (1)$$

The use of PCA scores can provide collinearity and noise-reduction benefits, but it requires the specification of the appropriate number of principal components (PCs). In this paper, two and three principal components were used for the UV-Vis (explained variance, R^2, equal to 97%) and NIR (R^2: 94%) spectra cluster analysis, respectively.

Furthermore, the nearest neighbor method was used to define a cluster [23]. Specifically, the distance between any two clusters was defined as the minimum of all pair-wise distances between object of each cluster; the two clusters with the minimum distance were then merged.

3. Results

3.1. Spectrophotometric Characterization of Starch–Triiodide Complex

Eighteen distinct flour samples were characterized through the measurement of absorption spectra of their starch–triiodide complex, and the obtained spectra are reported in Figure 1. These spectra show several bands within the investigated UV-Vis region.

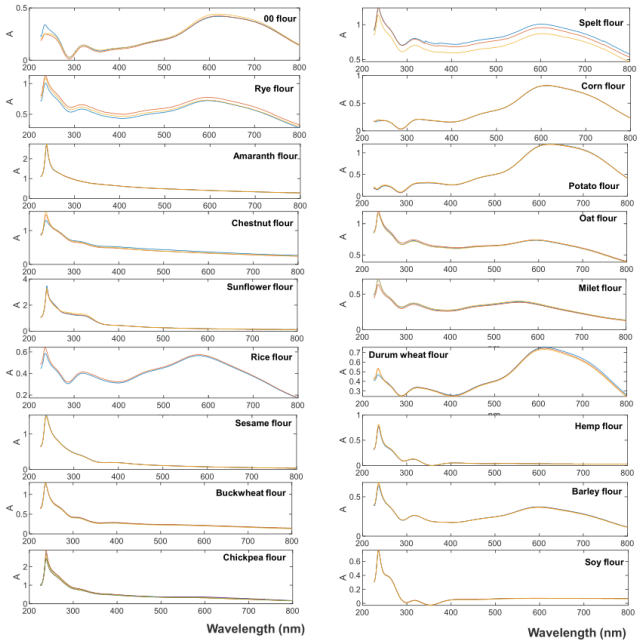


Figure 1. UV-Vis spectra of the eighteen investigated starch samples treated with a potassium triiodide solution 0.03 M to form the corresponding iodine complexes. T = 298 K, pH = 7.0. For each flour type, the different replicates, with different color lines, are reported.

In the UV region, three bands are observed at about 240 nm, 270 nm (broad shoulder) and 320 nm. Amylose and amylopectin do not show any signal in this spectral region; therefore, the signals observed are due to impurities or degradation products. In particular, the observed bands could be attributed, at least partly, to the presence in the flour of proteins or hydrolyzed amino acids which give well-defined signals in the spectral regions 220–240 nm (backbone) and 260–280 nm (aromatic amino acid residues). Phenolic compounds and ferulic acid were observed, respectively, at 280 nm and 320 nm in flour extracts [1] and might also be present in our samples. Specifically, the first peak at around 240 nm exhibits lower intensities in the corn and potato flours, while the second peak at approximately 320 nm is nearly absent in the amaranth flour.

Two additional peaks are observable in the range between 400 and 800 nm; they are evident for almost all flour samples except for amaranth, chestnut, sunflower, sesame, hemp and buckwheat flours. In particular, the first peak could be considered a shoulder peak of the second one with lower intensities between 400 and 500 nm. These two bands could be attributed to the absorption of amylopectin–iodide and amylose–iodide complexes [2,3]. The differences in their intensity may be due to the preliminary sample treatment, where starch is extracted in water, and amylose is indeed more soluble compared to amylopectin as well as to the botanical origin of the starch, which influences the ratio of amylose to amylopectin and their specific affinities for triiodide ion binding [3,4]. Upon visually inspecting the UV-Vis spectra obtained for various types of starch, differences are evident not only in the intensity of the bands but also in the wavelength values corresponding to the maximum absorbance of the amylose–iodide complex. Specifically, for 00 flour, the maximum peak occurs at 605 nm; for rye flour, it is at 595 nm; for rice flour, it is at 577 nm; for spelt flour, it is at 598 nm; for corn flour, it is at 609 nm; for potato flour, it is at 624 nm; for oat flour, it is at 599 nm; for millet flour, it is at 566 nm; for durum wheat flour, it is at 614 nm; and for barley flour, it is at 600 nm. As regards the intensity of the bands,

potato, spelt and durum wheat samples show the highest absorbance, while amaranth, sunflower, chickpea and buckwheat have the lowest ones. These observed variations could be attributed to several factors, such as inherent starch concentration, matrix effects from proteins and lipids, and variations in amylose/amylopectin ratios. In particular, the higher absorbance in potato, for instance, may be linked to its substantial starch content. Potato flour is essentially composed of starch, and a higher starch concentration can result in increased absorbance in the spectral measurements. The presence of proteins and lipids in flours, such as those in sunflower with a significant lipid fraction, can influence the extraction process. Therefore, this matrix effect may lead to variations in absorbance as different components in the flour matrix interact with the analytical method. Finally, differences in the amylose/amylopectin ratios among flours can impact the quantity of starch available for extraction. Flours with varying ratios may exhibit different interactions with the extraction process, affecting the measured absorbance.

Considering the richness of information held in the whole signals, all the obtained UV-Vis spectra were analyzed by principal component analysis. The UV-Vis spectra were organized into a dataset of 55 × 576 dimensions (samples and replicates on the raw × UV-Vis variables on the column) and mean centered. In particular, 576 columns correspond to the data points in each UV signal. In detail, the UV spectra were acquired over a wavelength range from 200 to 800 nm with a resolution of 1 nm. Each point in the UV signal represents the absorbance value at a specific wavelength, and the entire UV signal is constituted by these 576 absorbance values. The PCA model was developed using two principal components, according to their explained variance (R^2: 97%).

In Figure 2, the PC1 vs. PC2 scores plot is reported, representing the different flour samples with different symbols and colors. In the first principal component (PC1), the most significant differences emerge between sunflower, chickpea, and amaranth flours, which have positive PC1 values, and potato flour, which displays negative PC1 values. On the other hand, along the second principal component (PC2), spelt and potato with positive PC2 values are opposed to soy and hemp with negative values. From a comprehensive analysis of the figure, similarities emerge among the following groups: (i) soy and hemp (with negative scores for both components), (ii) corn, durum wheat, and rice, (iii) oat and rye, (iv) barley, millet, buckwheat, and Type 00 flour, and (v) amaranth and chickpeas.

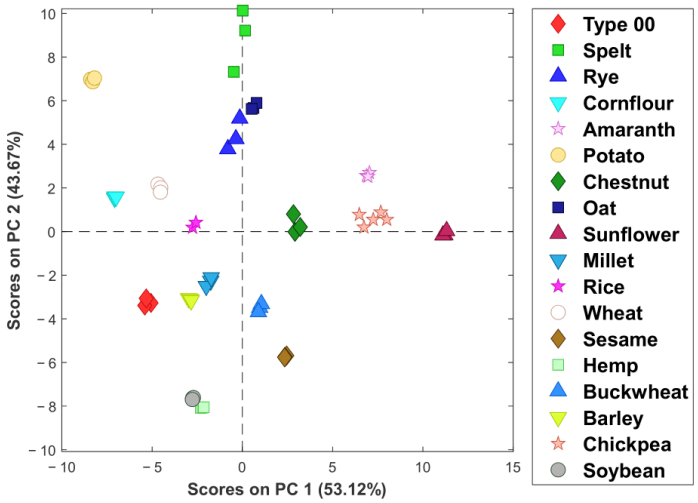
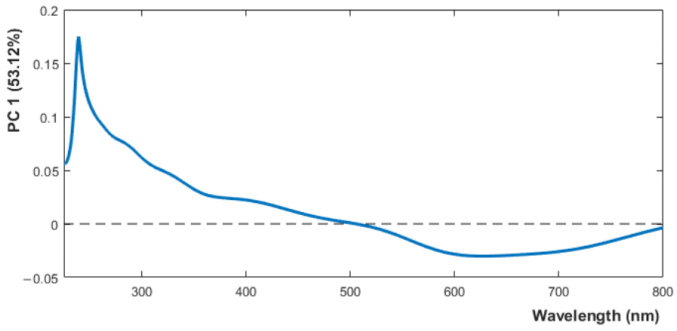
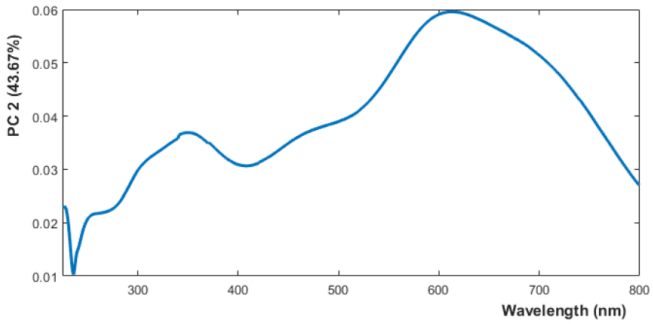


Figure 2. PC1 vs. PC2 scores plot obtained by the PCA analysis of UV-Vis spectra.

From the PC1 loadings figure (Figure 3a), it can be possible to point out the following significant regions in differentiating the samples on PC1: the range between 250 and 400 nm with positive values and the region corresponding to the absorption of the starch-triiodide complex between 500 and 700 nm with negative values. Specifically, flour samples with positive PC1 values (sunflower, chickpea and amaranth) seem to present higher intensities in the first part of the spectrum and lower intensities in the latter part.



(**a**)



(**b**)

Figure 3. Loadings plot of PC1 (**a**) and PC2 (**b**) of PCA applied on UV-Vis dataset.

In the second principal component (Figure 3b), the most influential area relates to the absorption of the starch–iodide complex (range from 400 to 700 nm) with positive values. Consequently, the samples of potatoes, spelt, oats, and rye exhibit higher intensities compared to soy and hemp.

In addition to the visual inspection of starch spectra and the rough differentiation from the scores plot of principal component analysis, all the UV-Vis spectra were examined by cluster analysis.

Figure 4 shows a dendrogram of hierarchical clustering of k-nearest neighbor distances (Section 2.6). In particular, it was obtained considering the UV spectra coming from all 18 samples (Session 2.2). Each flour was analyzed in triplicate except for chickpea flour, which featured five replicates. It is worth noting the uniqueness of many samples, namely potato, chestnut, sunflower, durum wheat, sesame, buckwheat, rice, corn, spelt and Type 00, that cannot be associated at any clusters when the distance values are lower than two. These differences can be due to several reasons such as the different amylose–amylopectin ratios, degrees of polymerization, helical structures, granule sizes and other physical and chemical properties of starch.

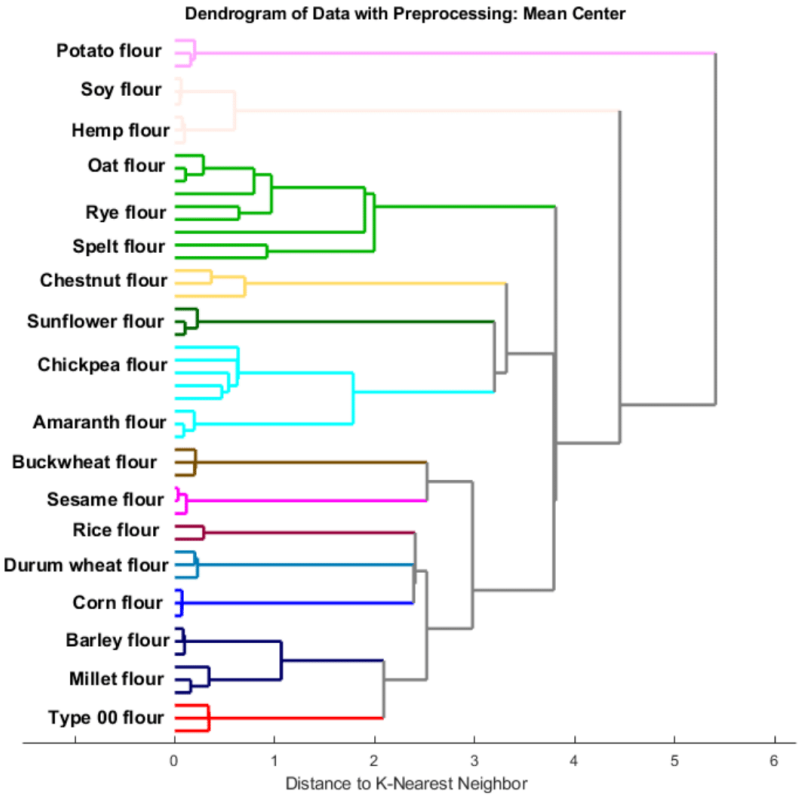


Figure 4. A dendrogram of cluster analysis of UV-Vis spectra of investigated starch samples.

On the other hand, it is possible to highlight a significant similarity among other samples, such as millet and barley, chickpea and amaranth, oat and rye, or soybean and hemp starches, which underscores the need to employ additional techniques, such as scanning electron microscopy, for a more selective differentiation among them.

3.2. Near Infrared Spectral Analysis

Although UV-Vis spectra have yielded valuable insights into assessing the similarities and differences between the samples, this method requires sample pre-treatment. To seek a simpler and faster alternative, the same flour samples were analyzed using NIR spectroscopy without any prior preparation.

The overall raw NIR spectra are graphically shown in Figure 5. All samples showed almost a similar trend in the shape of spectra except for hemp flour, presenting a different trend in the beginning and at the end of the respective baseline.

The spectra revealed prominent absorbance regions, particularly around 1200 nm, 1470 nm, 1580 nm, and 1665 nm. The absorption band at 1200 nm corresponded to the second overtone of C–H stretch associated with lipids. The substantial absorbance peaks at 1470 nm were linked to the first overtone of O–H stretching, which is indicative of the moisture content or starch [19,20,25]. The absorbance at 1580 nm was attributed to the first overtone of O–H stretching and was associated with starches, while the peak around 1665 nm was associated with the first overtone of C–H stretching and aromatic compounds [19]. These significant absorbance regions align with findings reported in the literature [19,20,25]. Taking into consideration the aims of this research, it was decided

to focus attention on spectral regions mainly influenced by the presence of water, since it could be potentially related to the different structures of starches of different botanical origin. In particular, the amount of water absorbed varies depending on the botanical species, genotype, and the degree of organization of the starch granules [27]. Furthermore, the characteristics of starch granules are highly influenced by the moisture content of the medium [27].

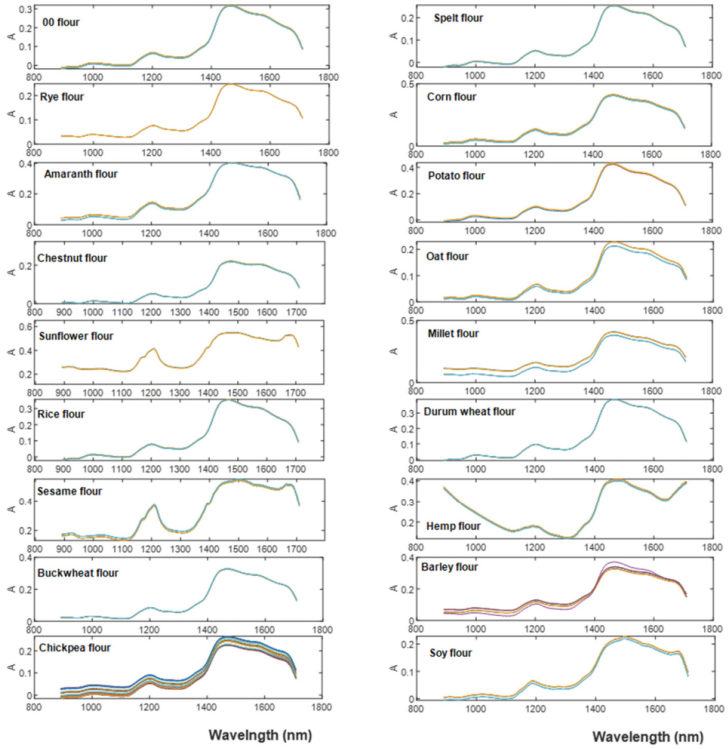


Figure 5. NIR spectra of the entire set of flour samples. For each flour type, the different replicates, with different color lines, are reported.

The spectra revealed prominent absorbance regions, particularly around 1200 nm, 1470 nm, 1580 nm, and 1665 nm. The absorption band at 1200 nm corresponded to the second overtone of C–H stretch associated with lipids. The substantial absorbance peaks at 1470 nm were linked to the first overtone of O–H stretching, which is indicative of the moisture content or starch [19,20,25]. The absorbance at 1580 nm was attributed to the first overtone of O–H stretching and was associated with starches, while the peak around 1665 nm was associated with the first overtone of C–H stretching and aromatic compounds [19]. These significant absorbance regions align with findings reported in the literature [19,20,25]. Taking into consideration the aims of this research, it was decided to focus attention on spectral regions mainly influenced by the presence of water, since it could be potentially related to the different structures of starches of different botanical origin. In particular, the amount of water absorbed varies depending on the botanical species, genotype, and the degree of organization of the starch granules [27]. Furthermore, the characteristics of starch granules are highly influenced by the moisture content of the medium [27].

Consequently, to consider the range between 1275 and 1600 nm could be an attempt to obtain information about the presence of a pattern among the different investigated starches according to their botanical origin. Therefore, PCA analysis was carried out only on this region; in particular, the spectral data were pre-treated using SNV before PCA analysis to reduce the multiplicative interferences of scatter and particle size of raw spectra [16] and mean centered. A PCA model was built considering three PCs explaining 98% of the total variance.

Separation between samples was clearly observed in the PCA scores plot of PC1 and PC2 (Figure 6).

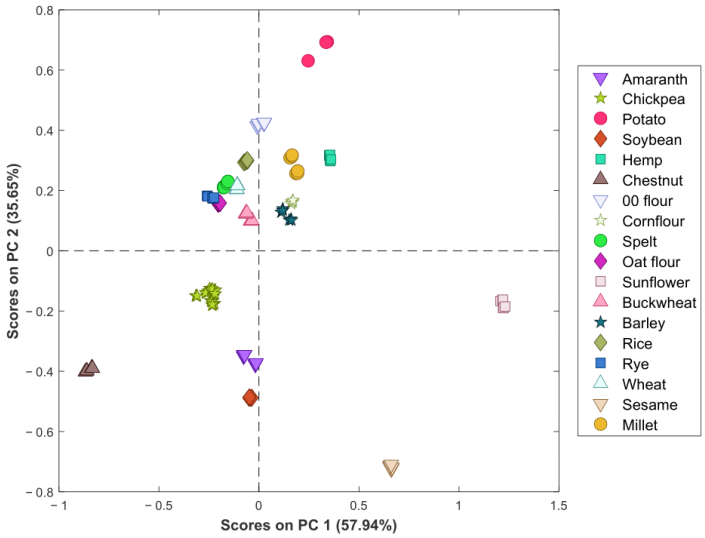


Figure 6. PC1 vs. PC2 scores plot obtained by the PCA analysis of NIR spectra in the range between 1275 and 1600 nm.

In particular, a distinct behavior is observed for sunflower and sesame flours (with positive values on PC1 and negative values on PC2), potato flours (with higher values on PC2) and chestnut flours (with more negative values on PC1). Furthermore, the second principal component distinguishes the samples of chickpea, amaranth, and soy, with negative values on PC2, from the other samples, which have positive values on PC2. These observations partially align with the results obtained from the UV-Vis analysis, especially concerning the differences observed between hemp and potato flours with soybean flour.

The PC1 and PC2 loadings plot gives information about the wavelengths that contributed to sample separation (Figures 7a and 7b, respectively).

The highest loadings observed at around 1400 and 1450 nm for PC1 and PC2 were related to water or starch as the first O-H stretching overtone. Upon a comprehensive analysis of both Figures 6 and 7, it is evident that sunflower and sesame flour, with positive PC1 and negative PC2 scores (fourth quadrant in Figure 6), are characterized by a notable increase in the intensity of the 1400 nm peak and a corresponding decrease in the intensity of the 1450 nm peak. Notably, the band around 1450 nm, as indicated in the literature, is associated with water or starch as the first O–H stretching overtone, while the peak at 1400 nm corresponds to the CH_2 stretching of lipids. Moreover, the positioning of potato flour, characterized by the highest PC2 score values, appears to be significantly influenced by elevated absorbance around 1450 nm. In contrast, chestnut flour, with negative PC2 scores, exhibits the opposite trend.

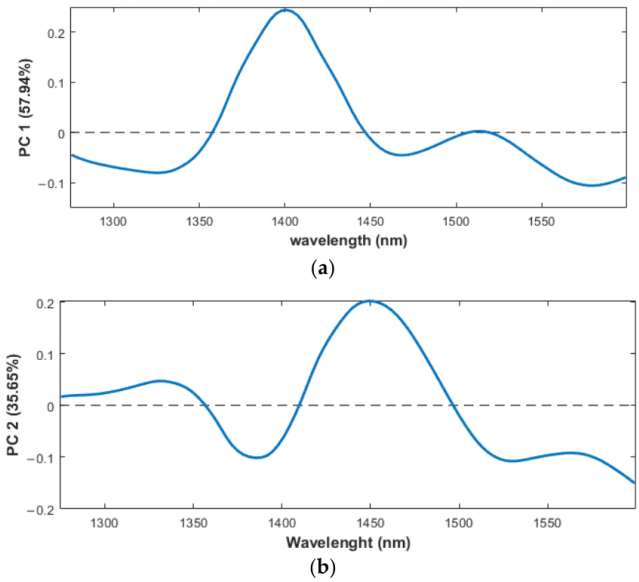


Figure 7. Loadings plot of PC1 (**a**) and PC2 (**b**) of PCA applied on NIR dataset.

The third principal component mainly distinguishes hemp and soybean flour mainly due to the contribution of the NIR band around 1500–1550 nm. The respective score and loading figures are reported as Figures S1 and S2 in Supplementary Materials.

The result of cluster analysis on NIR spectra in the region between 1275 and 1600 nm is reported in Figure 8. In this case, four replicates were acquired for each sample except for the control sample, chickpea flour, which presented 27 replicates (Session 2.4). The results display a significant consistency with what was found in the UV-Vis analysis.

In fact, a substantial similarity is observed between amaranth and chickpea flour, as well as between rye flour and oat flour, and between barley and millet flour. Furthermore, it can be observed that according to UV-Vis results, there are samples that are different from the others, such as sunflower, sesame, chestnut and potato flours.

3.3. Starch Morphology

Both UV-Vis and NIR spectroscopic analysis coupled with chemometrics show similarities among some investigated samples, such as barley and millet, rye and oats, and chickpeas and amaranth. This similarity complicates their differentiation based on botanic origin. Consequently, in these cases, the support of commonly used analytical techniques such as scanning electron microscopy is essential to obtain additional information.

Starch granules are microscopic in size, and their morphology varies between different shapes such as oval, ellipsoidal, spherical, smooth, angular, and lenticular, depending on their botanical origin. In amyloplasts, starch granules are present singly or in groups. Common cereals such as wheat, barley and rye contain two types of starch granules: type A, with lenticular shape and large size; type B with spherical shape and small size [28,29].

For obtaining a clear visualization of starch granules, their isolation from flour is essential due to the presence of fiber and starch protein cluster. However, in the present work, the aim was to investigate the possibility of obtaining information reducing as much as possible any pretreatment of sample. Therefore, the SEM images were directly acquired on raw flour after the coating of the sample surface with a thin layer of gold (Section 2.5) with the aim of characterizing the morphology of certain flours that appeared to be similar to others in statistical analysis.

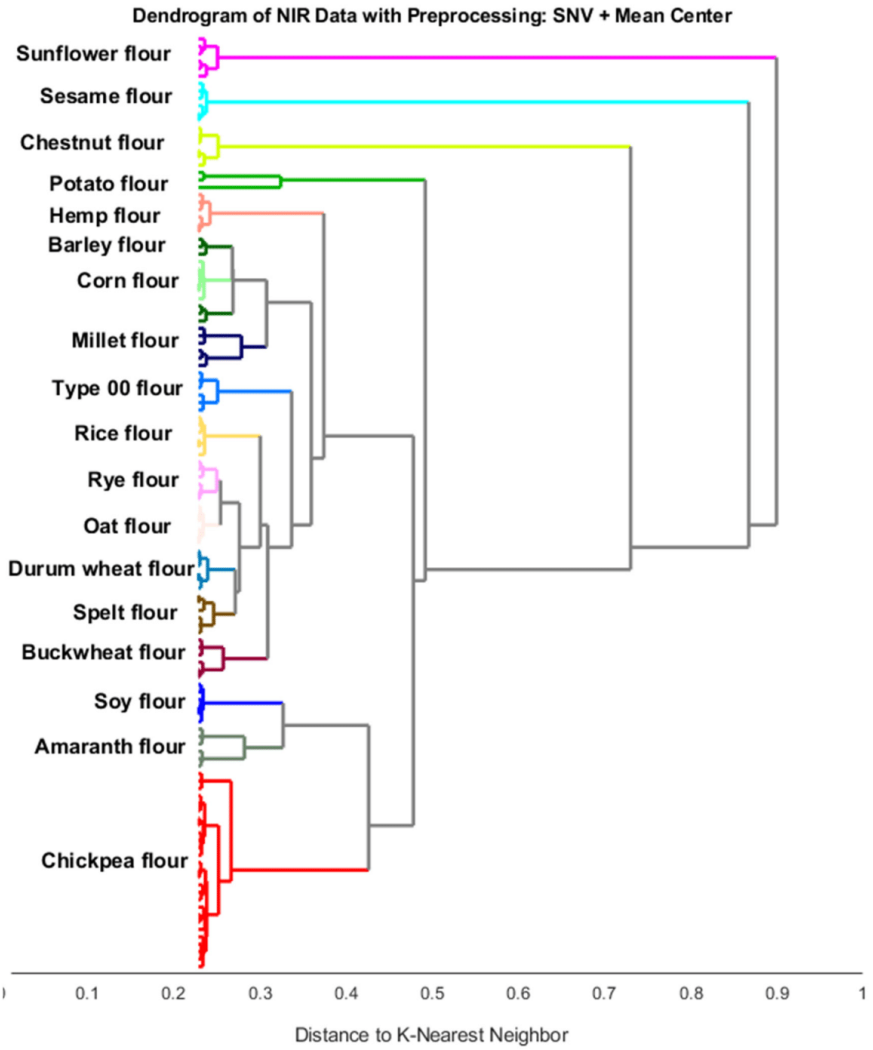
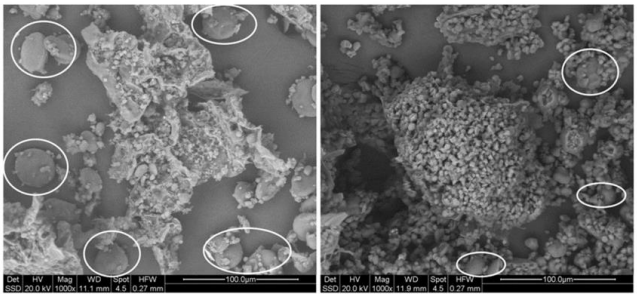


Figure 8. A dendrogram of cluster analysis of NIR spectra of investigated flour samples.

For instance, in Figure 9a,b, SEM images obtained on the samples of rye and oat flour were reported. The starch granules of rye flour (Figure 9a) have a lenticular shape (white circles in the figure) with a wrinkled surface and medium size (about 30 µm). On the surface of the starch granules, it can be also possible to see other smaller granules, which is probably due to the presence of associated proteins. In the case of oat flour (Figure 9b), a different structure is observed than in the previous samples. In this case, there are granules (white circles in the figure) with irregular shapes and sizes and rough surfaces, agglomerates consisting of small granules and irregular shapes, as well as the presence of proteins and cell walls.



(**a**) (**b**)

Figure 9. SEM image of (**a**) rye (highlighted with white circles) and (**b**) oat (highlighted with white circles) flours at HV 20.0 kV, Mag 1000×, Spot 4.5, HWF 0.27 mm.

The previous statistical analysis revealed that among the various flours, potato, rice, Type 00, and corn flours were found to be distinctive. These differences can be found in the structure of starch, as can be seen in Figures 10–13, respectively.

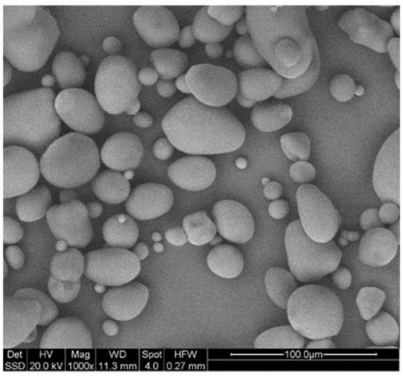


Figure 10. SEM image of potato flour at HV 20.0 kV, Mag 1000×, Spot 4.0, HWF 0.27 mm.

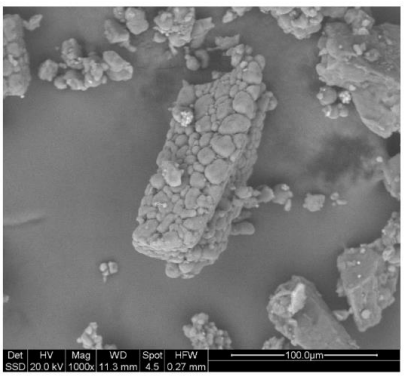


Figure 11. SEM image of rice flour at HV 20.0 kV, Mag 1000×, Spot 4.5, HWF 0.27 mm.

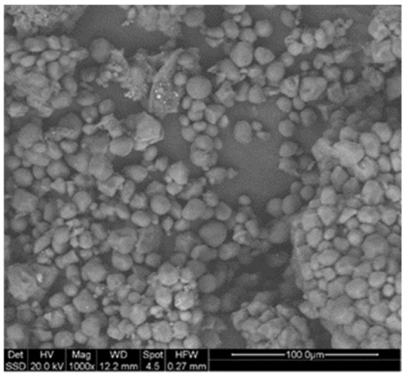


Figure 12. SEM image of corn flour at HV 20.0 kV, Mag 1000×, Spot 4.5, HWF 0.27 mm.

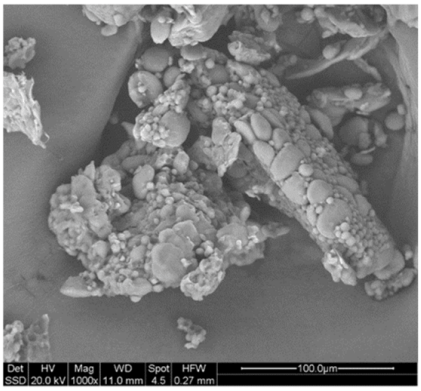


Figure 13. SEM image of Type 00 flour at HV 20.0 kV, Mag 1000×, Spot 4.5, HWF 0.27 mm.

The starch granules of potato flour (Figure 10) have smooth, polished surfaces with an oval shape. The smallest granules have a diameter of about 25 µm; larger granules reach up to about 60 µm. The starch granules of rice flour (Figure 11) are in the form of agglomerates consisting of granules of irregular shapes and sizes and smooth surfaces.

Examining the SEM image for Type 00 flour (Figure 12), an agglomerate consisting of starch granules with a lenticular shape of varying size and smooth surface is observed as well as protein. The starch granules of corn flour (Figure 13) are characterized by an angular shape, smooth surface and regular size of about 14 µm, and they associate to form agglomerates.

4. Conclusions

In this research, eighteen commercial flours, intended for food use, were chemically characterized by an untargeted approach, evaluating the possibility of developing rapid and alternative methods for identifying the botanical origin of starch. In particular, the used approach allowed exploring and analyzing a multitude of signals simultaneously, providing a more holistic and informative perspective on the characteristics of the examined samples. Specifically, all investigated samples were analyzed by two spectroscopic techniques, UV-Visible and near-infrared (NIR) spectroscopy; the signals obtained were considered as fingerprints of the investigated samples and analyzed by chemometric techniques. Considering the limited number of examined samples for each type of flour, it was not possible to develop classification models (which would have required a representative

sampling in terms of the number for each type of investigated flours). However, an initial exploratory investigation was carried out using principal component analysis and cluster analysis, which provided interesting preliminary information on the similarities and/or differences between starches of different botanical origins based on both UV-Vis and NIR acquired spectra.

In particular, the UV-Vis obtained model shows an interesting pattern among samples according to their botanical origin, distinguishing samples such as potato, chestnut, sunflower, durum wheat, sesame, buckwheat, rice, corn, spelt and Type 00 flours. However, some similarities were found for other flours, barley and millet, rye and oats, and chickpeas and amaranth, highlighting the need to use other techniques such as Scanning Electron Microscopy (SEM) to obtain supporting information.

Although this study is still a preliminary investigation and far from leading to the development of a classification model, the results obtained provide a solid premise for the development of an analytical approach based on the use of UV-Vis spectroscopy as a fingerprint technique for the identification of the botanical origin of starches. Regarding NIR spectroscopy, promising outcomes have emerged from the analysis of NIR regions, which are predominantly attributed to the presence of water and starch, suggesting its potential as an alternative methodology for distinguishing starches from various botanical sources without any sample pre-treatment.

Some remarks can be highlighted from the results obtained. The choice between NIR and UV-Vis spectroscopy for characterizing different flours depends on the challenges and goals of the research, as each technique offers distinct advantages and disadvantages. UV-Vis spectroscopy is particularly effective for starch analysis, as it provides specificity to differentiate starches from different botanical sources. However, one limitation is that UV-Vis often requires pretreatment of the sample, making it unusable after analysis. On the other hand, NIR spectroscopy has the advantage of not requiring pretreatment of the sample, allowing it to be recovered. However, NIR spectra can be affected by various compounds, such as the presence of water, which can overlap with the starch signal, leading to a non-selective technique. In future practical applications, the combination of both techniques could offer a powerful approach for characterizing flour. Leveraging the strengths of UV-Vis in starch specificity and NIR in non-destructive analysis will certainly provide a more complete and comprehensive characterization of flour. In fact, future research could be addressed in the use of mathematical–statistical models based on a data fusion approach [30] as well as in the increasing the number of investigated samples for each type of studied starch. However, investigation by scanning electron microscopy is certainly useful in supplementing the model results in case of overlapping information.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/chemosensors12010001/s1; Figure S1: Scores plot of PC3 of PCA applied on NIR dataset; Figure S2: Loadings plot of PC3 of PCA applied on NIR dataset.

Author Contributions: Conceptualization, M.C., C.D. and L.S.; Data curation, S.P., A.D. and L.S.; Formal analysis, S.P., A.D. and G.F.; Investigation, S.P., M.B., G.F. and L.S.; Methodology, S.P., M.B., M.C., C.D. and L.S.; Project administration, C.D.; Resources, C.D.; Software, M.C., A.D. and L.S.; Supervision, M.C. and C.D.; Validation, S.P., M.B., M.C. and A.D.; Visualization, A.D., G.F. and L.S.; Writing—original draft, S.P., M.B. and C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Acknowledgments: The authors would like to thank Mauro Zapparoli (Centro Interdipartimentale Grandi Strumenti, University of Modena and Reggio Emilia, Italy) for his technical support in SEM images acquisition. Furthermore, the authors would like to express their deepest gratitude to Quantum Design Italy (https://qdindustria.it/, accessed on 1 December 2023) that provided a poliSPEC NIRe portable instrument to perform NIR measurements.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Santana, Á.L.; Meireles, M.A.A. New Starches Are the Trend for Industry Applications: A Review. *Food Public Health* **2014**, *4*, 229–241. [CrossRef]
2. Cheng, H.; Chen, L.; McClements, D.J.; Yang, T.; Zhang, Z.; Ren, F.; Miao, M.; Tian, Y.; Jin, Z. Starch-Based Biodegradable Packaging Materials: A Review of Their Preparation, Characterization and Diverse Applications in the Food Industry. *Trends Food Sci. Technol.* **2021**, *114*, 70–82. [CrossRef]
3. Rodrigues, A.; Emeje, M. Recent Applications of Starch Derivatives in Nanodrug Delivery. *Carbohydr. Polym.* **2012**, *87*, 987–994. [CrossRef]
4. dos Santos Alves, M.J.; Chacon, W.D.C.; Gagliardi, T.R.; Agudelo Henao, A.C.; Monteiro, A.R.; Ayala Valencia, G. Food Applications of Starch Nanomaterials: A Review. *Starch Stärke* **2021**, *73*, 2100046. [CrossRef]
5. Shevkani, K.; Singh, N.; Bajaj, R.; Kaur, A. Wheat Starch Production, Structure, Functionality and Applications—A Review. *Int. J. Food Sci. Technol.* **2017**, *52*, 38–58. [CrossRef]
6. Garcia, M.A.V.T.; Garcia, C.F.; Faraco, A.A.G. Pharmaceutical and Biomedical Applications of Native and Modified Starch: A Review. *Starch Stärke* **2020**, *72*, 1900270. [CrossRef]
7. Copeland, L.; Blazek, J.; Salman, H.; Tang, M.C. Form and Functionality of Starch. *Food Hydrocoll.* **2009**, *23*, 1527–1534. [CrossRef]
8. Pérez, S.; Bertoft, E. The Molecular Structures of Starch Components and Their Contribution to the Architecture of Starch Granules: A Comprehensive Review. *Starch/Staerke* **2010**, *62*, 389–420. [CrossRef]
9. Liu, Q. Understanding Starches and Their Role in Foods. In *Food Carbohydrates: Chemistry, Physical Properties and Applications*; Routledge: London, UK, 2005; p. 340. ISBN 978-0-8493-1574-9.
10. Peris-Tortajada, M. Chapter 6—Measuring Starch in Food. In *Starch in Food*, 2nd ed.; Woodhead Publishing Series in Food Science, Technology and Nutrition; Sjöö, M., Nilsson, L., Eds.; Woodhead Publishing: Sawston, UK, 2018; pp. 255–281, ISBN 978-0-08-100868-3.
11. Sakač, N.; Karnaš, M.; Dobša, J.; Jozanović, M.; Gvozdić, V.; Kovač-Andrić, E.; Kraševac Sakač, M.; Šarkanj, B. Application of Spectrophotometric Fingerprint in Cluster Analysis for Starch Origin Determination. *Food Technol. Biotechnol.* **2020**, *58*, 5–11. [CrossRef]
12. Analysis of Starch in Food Systems by High-Performance Size Exclusion Chromatography—Ovando-Martínez—2013—Journal of Food Science—Wiley Online Library. Available online: https://ift.onlinelibrary.wiley.com/doi/10.1111/1750-3841.12037 (accessed on 13 November 2023).
13. Alcázar Alay, S.C.; Meireles, M.A. Physicochemical Properties, Modifications and Applications of Starches from Different Botanical Sources. *Food Sci. Technol.* **2015**, *35*, 215–236. [CrossRef]
14. Chatel, S.; Voirin, A.; Artaud, J. Starch Identification and Determination in Sweetened Fruit Preparations. 2. Optimization of Dialysis and Gelatinization Steps, Infrared Identification of Starch Chemical Modifications. *J. Agric. Food Chem.* **1997**, *45*, 425–430. [CrossRef]
15. Liu, Y.; Chao, C.; Yu, J.; Wang, S.; Wang, S.; Copeland, L. New Insights into Starch Gelatinization by High Pressure: Comparison with Heat-Gelatinization. *Food Chem.* **2020**, *318*, 126493. [CrossRef] [PubMed]
16. Chen, L.; Zhang, H.; McClements, D.J.; Zhang, Z.; Zhang, R.; Jin, Z.; Tian, Y. Effect of Dietary Fibers on the Structure and Digestibility of Fried Potato Starch: A Comparison of Pullulan and Pectin. *Carbohydr. Polym.* **2019**, *215*, 47–57. [CrossRef] [PubMed]
17. Yoon, J.-W.; Jung, J.-Y.; Chung, H.-J.; Kim, M.-R.; Kim, C.-W.; Lim, S.-T. Identification of Botanical Origin of Starches by SDS-PAGE Analysis of Starch Granule-Associated Proteins. *J. Cereal Sci.* **2010**, *52*, 321–326. [CrossRef]
18. Ndlovu, P.F.; Magwaza, L.S.; Tesfay, S.Z.; Mphahlele, R.R. Vis-NIR Spectroscopic and Chemometric Models for Detecting Contamination of Premium Green Banana Flour with Wheat by Quantifying Resistant Starch Content. *J. Food Compos. Anal.* **2021**, *102*, 104035. [CrossRef]
19. Handbook of Near-Infrared Analysis. Available online: https://www.routledgehandbooks.com/doi/10.1201/b22513 (accessed on 13 November 2023).
20. Blanco, M.; Villarroya, I. NIR Spectroscopy: A Rapid-Response Analytical Tool. *TrAC Trends Anal. Chem.* **2002**, *21*, 240–250. [CrossRef]
21. Vitelli, M.; Mehrtash, H.; Assatory, A.; Tabtabaei, S.; Legge, R.L.; Rajabzadeh, A.R. Rapid and Non-Destructive Determination of Protein and Starch Content in Agricultural Powders Using near-Infrared and Fluorescence Spectroscopy, and Data Fusion. *Powder Technol.* **2021**, *381*, 620–631. [CrossRef]

22. Li Vigni, M.; Durante, C.; Cocchi, M. Chapter 3—Exploratory Data Analysis. In *Data Handling in Science and Technology*; Chemometrics in Food Chemistry; Marini, F., Ed.; Elsevier: Amsterdam, The Netherlands, 2013; Volume 28, pp. 55–126.
23. Vandeginste, B.M.G.; Massart, D.; Buydens, L.; De Jong, S.; Lewi, P.; Verbeke, J. Chapter 30-Cluster Analysis. In *Data Handling in Science and Technology*; Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., Eds.; Elsevier: Amsterdam, The Netherlands, 1998; Volume 20, Part 2; pp. 57–86. ISBN 9780444828538.
24. Giussani, B.; Gorla, G.; Riu, J. Analytical Chemistry Strategies in the Use of Miniaturised NIR Instruments: An Overview. *Crit. Rev. Anal. Chem.* **2022**, 1–33. [CrossRef]
25. Zhao, H.; Guo, B.; Wei, Y.; Zhang, B. Effects of Grown Origin, Genotype, Harvest Year, and Their Interactions of Wheat Kernels on near Infrared Spectral Fingerprints for Geographical Traceability. *Food Chem.* **2014**, *152*, 316–322. [CrossRef]
26. Stuart, B. Spectral Analysis. In *Infrared Spectroscopy: Fundamentals and Applications*; John Wiley & Sons Ltd.: Chichester, UK, 2004; pp. 45–70. ISBN 978-0-470-854273.
27. Kovrlija, R.; Goubin, E.; Rondeau-Mouro, C. TD-NMR Studies of Starches from Different Botanical Origins: Hydrothermal and Storage Effects. *Food Chem.* **2020**, *308*, 125675. [CrossRef]
28. Pérez, S.; Baldwin, P.M.; Gallant, D.J. Chapter 5—Structural Features of Starch Granules I. In *Starch*, 3rd ed.; Food Science and Technology; BeMiller, J., Whistler, R., Eds.; Academic Press: San Diego, CA, USA, 2009; pp. 149–192. ISBN 978-0-12-746275-2.
29. de Miranda, J.A.T.; de Carvalho, L.M.J.; de Castro, I.M.; de Carvalho, J.L.V.; de Alcântara Guimarães, A.L.; de Macêdo Vieira, A.C. Scanning Electron Microscopy and Crystallinity of Starches Granules from Cowpea, Black and Carioca Beans in Raw and Cooked Forms. *Food Sci. Technol* **2019**, *39*, 718–724. [CrossRef]
30. Silvestri, M.; Bertacchini, L.; Durante, C.; Marchetti, A.; Salvatore, E.; Cocchi, M. Application of Data Fusion Techniques to Direct Geographical Traceability Indicators. *Anal. Chim. Acta* **2013**, *769*, 1–9. [CrossRef] [PubMed]

foods

MDPI

Review

How Chemometrics Can Fight Milk Adulteration

Silvia Grassi [1], Maria Tarapoulouzi [2], Alessandro D'Alessandro [3], Sofia Agriopoulou [4], Lorenzo Strani [3,*] and Theodoros Varzakas [4,*]

[1] Department of Food, Environmental and Nutritional Sciences (DeFENS), Università degli Studi di Milano, Via Celoria, 2, 20133 Milano, Italy
[2] Department of Chemistry, Faculty of Pure and Applied Science, University of Cyprus, P.O. Box 20537, Nicosia CY-1678, Cyprus
[3] Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Via Campi 103, 41125 Modena, Italy
[4] Department of Food Science and Technology, University of the Peloponnese, Antikalamos, 24100 Kalamata, Greece
* Correspondence: lostrani@unimore.it (L.S.); t.varzakas@uop.gr (T.V.)

Abstract: Adulteration and fraud are amongst the wrong practices followed nowadays due to the attitude of some people to gain more money or their tendency to mislead consumers. Obviously, the industry follows stringent controls and methodologies in order to protect consumers as well as the origin of the food products, and investment in these technologies is highly critical. In this context, chemometric techniques proved to be very efficient in detecting and even quantifying the number of substances used as adulterants. The extraction of relevant information from different kinds of data is a crucial feature to achieve this aim. However, these techniques are not always used properly. In fact, training is important along with investment in these technologies in order to cope effectively and not only reduce fraud but also advertise the geographical origin of the various food and drink products. The aim of this paper is to present an overview of the different chemometric techniques (from clustering to classification and regression applied to several analytical data) along with spectroscopy, chromatography, electrochemical sensors, and other on-site detection devices in the battle against milk adulteration. Moreover, the steps which should be followed to develop a chemometric model to face adulteration issues are carefully presented with the required critical discussion.

Keywords: fraud; authentication; dairy; clustering; classification; regression; validation

Citation: Grassi, S.; Tarapoulouzi, M.; D'Alessandro, A.; Agriopoulou, S.; Strani, L.; Varzakas, T. How Chemometrics Can Fight Milk Adulteration. *Foods* **2023**, *12*, 139. https://doi.org/10.3390/foods12010139

Academic Editor: Gianfranco Picone

Received: 26 November 2022
Revised: 10 December 2022
Accepted: 22 December 2022
Published: 27 December 2022

1. Introduction

Milk and milk products provide the human body with valuable nutritional components such as proteins, carbohydrates, vitamins, minerals, organic acids, and fat [1,2]. Milk's high protein content has attracted many consumers, making it a popular nutritional commodity [3]. The increasing consumption of milk and dairy products leads to many cases of adulteration [4,5]. A range of possible milk adulterants is described by Nascimento et al. [4].

The prices of milk differ primarily depending on the type of animal from which they come, whereas its availability is significantly affected by the season. These two factors are enough to cause problems in its market, as practices of replacing it with cheaper milk are common [6]. Goat's milk shows a nutritional profile superior to that of cows, as a result of which it is a priority for consumers not only in traditional dairy products such as cheese and yogurt, but also in liquid form. Its low production combined with its beneficial nutritional content makes this category of milk an attractive target for adulteration. Goat's milk is easily mixed with water, whey as well as cow's milk which is much cheaper. The latest fraud is increasingly worrying people because of their sensitivity to lactose and the allergic disorders that can be caused by cow's milk proteins [7]. An equally important adulteration

is related to the substitution of goat's milk with sheep's milk. In this case, the lower price of goat's milk compared to sheep's milk pushes the producers to this adulteration [6].

Fraud in milk production is carried out by admixture or substitution of inferior substances and sometimes dangerous products. The economically motivated adulteration (EMA) is the most important, aiming to gain profit by the addition of extraneous water, glucose or other sugars, non-dairy proteins such as soybean and pea protein isolates [8], various substances such as melamine, urea, maltodextrin, cheese whey (a byproduct of cheese production) [9], hypochlorite, dichromate, salicylic acid [10], and reconstituted milk powders to correct protein and/or density values [11]. A famous case of adulteration was recorded in China in 2013 when the substance melamine was detected in milk powder in infant milk products, which was added to increase the apparent protein content, with dramatic consequences for public health [12].

The deliberate addition of formaldehyde to raw milk is also illegal and considered a major adulteration, which aims to increase the shelf life of milk at room temperature. High moisture content is responsible for the rapid spoilage of milk. Therefore, formaldehyde provides preservative and antiseptic properties, and the ability to improve the appearance including the smell of milk. Furthermore, formaldehyde is toxic at low concentrations and is classified as a human carcinogen by the International Agency for Research on Cancer (IARC) [12,13].

Another form of adulteration is the replacement of milk fat with vegetable fats of lower economic value [14]. Among others, soybean oil has been mentioned in the adulteration of milk [15]. In addition, the recent EU regulations for foods designated as PDO (protected designation of origin), PGI (protected geographical indication), and TSG (traditional specialty guaranteed) require the inclusion on the label of the geographical origin of food. In the case of dairy products such as cheeses produced in a defined area with specific physicochemical and sensorial features, their geographical origin is put forward as an important indication [16].

Chemometrics plays a dominant role in the field of food adulteration as it relates a multitude of chemical analytical characteristics to the qualitative and quantitative analysis of food [17]. Deriving a fingerprint of each sample and reflecting its complex chemical composition could be a way to solve such difficult analytical tasks. Then, chemometric techniques can be used to develop classification models to classify samples into authentic/adulterated ones, or regression models aiming at quantifying a specific adulterant [8,18–21]. In this direction, both specific and non-specific fingerprinting can be implemented. Specific chemical analysis is based on the detection of organic species, mainly achieved by chromatographic techniques.

The non-specific fingerprinting approach relies on the implementation of instrumental methods to obtain a multivariate description of the chemical composition of the sample. These non-specific fingerprints can be obtained by different methodologies such as Fourier transform infrared spectroscopy (FT-IR), mid-infrared spectroscopy (MIR), Raman spectrometry, nuclear magnetic resonance (NMR), or mass spectrometry [22]. All these methodologies have been used in studies, which are relevant to authenticity and chemometrics in milk and dairy products [23–25]. In addition, near-infrared (NIR) spectroscopy has been used by several researchers to detect various forms of adulteration in both cow's milk and cow's milk products [26–28].

Vibrational spectroscopic techniques are rapid, low-cost, and non-destructive tests that require only limited training for processing. Results are evaluated using chemometric models to extract meaningful information that distinguishes different and significant groups by removing redundant data [29].

Data processing can be completed by principal component analysis (PCA) since it is amongst the most fundamental methods for multivariate data exploration [18]. PCA has been used along with other methodologies to help to differentiate fresh milk and reconstituted skim milk powder samples [11].

kNN (k-nearest neighbor), PLS-DA (partial least squares-discriminant analysis), and SIMCA (soft independent modeling of class analogy) are the most popular classification methods [30]. kNN and PLS-DA have been used for the detection of various types of adulteration, such as water, urea, cow's whey, and cow's milk in goat's milk samples [31]. SIMCA could also be employed to model the class of fresh types of milk. When addressing a specific adulterant quantification, the goal could be achieved by means of partial least squares (PLS) regression analysis, as demonstrated for the prediction of fresh milk adulteration with reconstituted skim milk powders [11].

Finally, in order to validate a chemometric approach, a sampling strategy should be followed taking into account the size and the representativeness of the sample along with intrinsic variability [32]. Sampling is closely associated with robustness and reliability. Other key parameters of authenticity and fraud not to be ignored are the heterogeneity of a food matrix and the presence of an undeclared substance to the geographical origin discrimination.

In this framework, the aim of this work is to give an overview of the recent application of different chemometric techniques—from clustering to classification and regression applied—to several analytical data—encompassing spectroscopy, chromatography, and electrochemical sensors—to fight milk adulteration. Further, a critical discussion is presented to schematize the steps which should be followed to develop a chemometric model to face adulteration issues.

2. Chemometric Approaches

2.1. Clustering

The definition of "cluster analysis" or "clustering" encompasses the techniques which split a set of samples (observations) into several groups or clusters. The outcome is usually represented as a vector of data, or a point (scatter) in a multidimensional space [33]. Clustering falls in the general category of unsupervised pattern recognition and numerical and mathematical taxonomy [33,34]. Natural grouping of data takes place based on some inherent similarity, as clustering is performed without any group labels, and this justifies the unsupervised pattern recognition [33,35]. Furthermore, it takes place based on similarities of the samples within the same group and others in different groups. Therefore, homogeneity is dominant within the same groups [34]. In practice, the most common approach to define similarity is the distance among the patterns; by lowering the distance (e.g., Euclidean distance which is a well-used dissimilarity measure) between the two objects, higher similarity and vice versa will be obtained [35,36].

Clustering is a valuable component of data analysis or machine learning-based applications such as regression, prediction, data mining, etc. [35]. Saxena et al. (2017) [35] stated that there are various ways to categorize clustering methods because it is difficult to define a cluster. In their paper, they suggested division into two different groups such as hierarchical and partitioning techniques, or in three categories based on application, density-based methods, model-based methods, and grid-based methods.

Hierarchical methods initially group the objects into small clusters of some samples, and these are next grouped into larger clusters, thus a dendrogram is produced, which is a tree-based depiction of each observation [36]. Optimization- partitioning methods split the samples into a few groups to optimize a particular feature e.g., total within-group distances. In this category, algorithms like k-*means clustering*, *Fuzzy* c-*means clustering*, etc., are included [33–35]. Density-based clustering is focused on the probability that data objects are drawn from a specific probability distribution and the overall distribution of the data is assumed to be a mixture of several distributions. Data points can be derived from different types of density functions (e.g., multivariate Gaussian or *t*-distribution), or from the same families but with different parameters. Model-based clustering works by detecting feature details for each cluster, where each cluster represents a concept or class. Decision trees and neural networks are the two most frequently used methods in this category. Grid-based clustering divides the space into a finite number of cells that make a grid structure on which all the operations for clustering are performed [35].

Recently, many evaluation criteria have been developed, and these are internal and external. Internal quality parameters include the sum of squared error, scatter criteria, Condorcet's criterion, the C-criterion, category utility metrics, and edge cut metrics. External quality criteria are related to the mutual information-based measure, Rand index, F-measure, Jaccard index, Fowlkes–Mallows index, and confusion matrix [35].

Clustering is applied to perform data reduction or compression for handling huge loads of data. It helps in compressing data information by grouping them into different sets of clusters. This helps us to choose what is useful or not by saving time from data processing along with data reduction [35]. Other uses contain data mining, document retrieval, image segmentation, and pattern classification [33].

In order to explore the use and development of clustering methods recently, Table 1 has been prepared to summarize the studies related to milk adulteration and authenticity.

Table 1. Recent studies (2015–2021) related to milk adulteration and authenticity in combination with clustering analysis.

Type of Milk	Target	Analytical Method(s)	Clustering Method	Approach	Reference
Milk adulteration					
Cow's, sheep's, and water buffalo's origin milk	Adulteration from different species' origin milk	FTIR	HCA	method	[37]
Bovine milk	Adulteration with urea	EIS	HCA	Euclidean distance	[36]
UTH milk samples (skimmed and semi-skimmed) and raw milk	Adulteration with cheese whey, based on quantification of caseinomacropeptide	FTIR-ATR	HCA	Euclidean distance and Ward's method	[38]
Cow milk	Adulteration with melamine and urea	Electrochemical biosensor	HCA	Ward's method	[39]
Bovine milk	Adulteration with formaldehyde, based on aldehydes and ketones	Colorimetric sensor array	HCA	-	[40]
UHT whole bovine milk and UHT goat milk	Adulteration with soymilk in bovine and goat milk, as well as bovine milk in goat milk.	NMR	CA	The minimum distance method	[41]
Raw cow milk	Adulteration with Sodium Salicylate, Dextrose, Hydrogen Peroxide, Ammonium Sulphate	Sensor system	k-means clustering algorithm	-	[42]
Milk authentication					
Powder and liquid milk	Type of milk based on metal profiles	ICP-OES	HCA	Euclidean distance and Ward's method	[43]
Organic and conventional milk	Type of milk (organic vs. conventional) based on organic status and trace element content	ICP-MS	HCA	Euclidean distance and Ward's method	[44]
Malaysian vs. milk from other countries	Geographical origin, based on metal content	ICP-MS	HCA	Ward's method	[45]

Table 1. *Cont.*

Type of Milk	Target	Analytical Method(s)	Clustering Method	Approach	Reference
-	Geographical origin, isotope ratios, metals, and fatty acids	CF-IRMS (δ ^{18}O), EA-IRMS (δ ^{13}C and δ^{15}N), GC (fatty acids), ICP-OES (Na, K, Mn, P, Zn, Ca, Fe, and Mg), and ICP-MS (other metals)	HCA	-	[46]
Cow milk	Geographical origin, based on stable isotope ratios	IRMS and CRDS	HCA	-	[47]
Raw milk	Geographical origin, based on stable isotope ratios and metal content	IRMS and ICP-MS	HCA and k-means clustering algorithm	HCA: Euclidean distance and Ward's method K means: 200 iterations and 25 random starting points	[48]
Cow, goat, camel, donkey, and yak milk	Species recognition based on sn-2 and sn-1,3 fatty acid composition and sterols	GC, GC-MS	HCA	-	[49]
Fresh buffalo, bovine, and donkey milk as well as processed milk samples (pasteurized and dried skimmed powder)	Species recognition based on amino acids, non-amino acids, and citric acid cycle metabolites	GC-MS	HCA	Euclidean distance and Ward's method	[50]
Reconstituted milk vs. UHT milk	Different content of peptides, lipids, and nucleic acids	UPLC–Q-TOF-MS combined with UPLC–MS/MS	HCA	-	[51]
Cow milk	Fat globule characteristics (diameter, membrane surface, and yield), fat, protein, fatty acids, calcium content	IR (fat, protein, and lactose contents), GC (fatty acids composition), atomic absorption spectrophotometry (calcium content)	HCA	Euclidean distance	[52]
Cow, goat, buffalo, and camel milk	Different seasons of milk collection, based on sterols in milk fat of different species' origin of milk	GC–MS-SIM	HCA	Euclidean distance	[53]

Abbreviations: CA = cluster analysis, CF-IRMS = continuous flow-isotope ratio mass spectrometer, CRDS = cavity ring-down spectroscopy, EA-IRMS = element analysis-isotope mass spectrometry, EIS = electrochemical impedance spectroscopy, FCM = fuzzy c-means, FTIR-ATR = Fourier transform infrared-attenuated total reflection, FTIR = Fourier transform infrared spectroscopy, GC = gas chromatography, GC-MS = gas chromatography-mass spectrometry, GC-MS-SIM = gas chromatography-mass spectrometry-single ion monitoring mode, HCA = hierarchical cluster analysis, ICP-MS = inductively coupled plasma mass spectrometry, ICP-OES = inductively coupled plasma emission spectroscopy, IR = infrared, IRMS = isotopic ratio mass spectrometry, UHT = ultra-high temperature, UPLC–MS/MS = UPLC–tandem mass spectrometry, UPLC–Q-TOF-MS = ultra-high performance liquid chromatography-quadrupole time-of-flight mass spectrometry.

Regarding milk adulteration studies, Cirak et al. [37] focused on determining milk species adulteration by using FTIR. HCA was conducted based on Ward's algorithm after having calculated the initial derivate by using a standard method. The produced 2D-dendrogram indicated that the types of origins (sheep, cow, and water buffalo origin, and adulterated samples in binary mixtures) were clustered correctly. Minetto et al. [36] applied HCA to detect urea in raw bovine milk samples, and the Euclidean distance was used to build the dendrogram. HCA helped them to find the more appropriate number of clusters which was used later in the classification of the samples. Vinciguerra et al. [38] used HCA as an exploratory treatment on the pre-processed measurements obtained by FTIR-ATR. By using both the Euclidean distance and Ward's method, a dendrogram was generated, however no pattern related to the caseinomacropeptide concentration was observed in the dendrogram, and multivariate regression was followed. Qualitatively, the adulterated groups with caseinomacropeptide were separated correctly in 3 groups: raw milk, skimmed milk, and semi-skimmed milk. Adulteration with melamine and urea in cow's milk was also studied by Ezhilan et al. [39], who developed an electrochemical biosensor to detect the two adulterants simultaneously. HCA application was useful to study the interrelationship of the factors affecting the model for measurements taken by using various combinations of concentrations of the adulterants. Mostafapour et al. (2021) [40] used a colorimetric array device. The authors commented that even if there are differences in the colorimetric schemes of the analytes, it is not a proper manner to group the samples after visual examination, thus chemometrics is used to perform the clustering. The HCA dendrograms showed highly accurate clustering of the studied carbonyl compounds, particularly eight different aldehydes and ketones. In addition, HCA showed that one sample from formaldehyde and one sample from acetophenone has been misclassified. Li et al. [41] used NMR to detect the metabolites as markers of different milk types. Clustering analysis (CA) was very useful as it provided similarities for the same species of milk as well as variations in different milk species by applying the minimum distance method. CA also separated the three milk types and showed that NMR and metabolites can differentiate these milk products. Sowmya et al. [42] during the pre-processing steps applied cluster analysis, i.e., the k-means clustering algorithm. The algorithm proceeded by calculating the centroid point of the dataset and the groups' mean points to build the new groups required. The aim was to see the grouping of samples, to identify the similarities in the same categories, and to check if the adulterants can be clustered by using raw spectra. Intraclass variation was performed.

Regarding milk authenticity, Souza et al. [43] studied the metal profile of powder and liquid milk samples to differentiate them based on the type of milk. HCA successfully confirmed the initial outcome of PCA, and it allows the visualization of a sample's trend to form two groups. Whole cow powder milk, whole goat powder milk, skimmed cow powder milk, and milk compounds powder fell in the first group due to their similar composition. A sample from the last group clustered at a longer distance from its group due to the high content of Zn. The second group consisted of whole and skimmed cow liquid milk and some yogurts. Rodriguez-Bermudez et al. [44] by applying HCA revealed a correct clustering based on the type of milk, organic vs. conventional. It was obvious that the variables (metal content) in both the organic and conventional sets were distinct. To determine the geographical origin, Zain et al. [45] measured the metal content of milk samples and due to different environmental conditions, and the samples clustered successfully by HCA. Ca, Na, Fe, Zn, Mn, K, Ba, and Mg are the metals that were significant for the samples' grouping regarding geographical origin. Xu et al. [46] worked also in terms of geographical origin by measuring isotope ratios, metals, and fatty acids and then by applying HCA. $\delta^{18}O$ measurements were taken by having the milk in the fluid state, but for $\delta^{13}C$, $\delta^{15}N$, and elemental and fatty acid measurements lyophilization took place. HCA aided to picture the correlation between the sample and each variable as HCA heatmaps were created. In addition, geographical origin was the target of Amenzou et al. [47], who studied the $^{13}C/^{12}C$, $^{15}N/^{14}N$, $^{18}O/^{16}O$. The application of HCA was very important to visualize the

samples in 3 important clusters. The stable isotope ratios analysis in combination with chemometrics showed a very good capability to indicate the geographical origin of milk. In a similar study, Podkolzin and Solovev [48] used HCA and the k-mean clustering algorithm and both methods showed an equal number of clusters with almost the same content. Karrar et al. [49] used HCA to evaluate the similarity in terms of sn-2 and sn-3 fatty acids in different milk-origin samples. HCA heatmaps were produced to present the content of sn-2 and sn-3 fatty acids in the samples. Bhumireddy et al. [50] applied HCA to group the samples based on intrinsic similarities in their GC-MS measurements. HCA heatmaps were produced using the log-transformed and normalized values of the relative abundance of 17 amino acids, and their high and low expressions in each sample were presented with different colors. Tan et al. [51] employed HCA to proceed to the clustering of the different biomarkers (peptides, lipids, and nucleic acids) and to demonstrate the chemical properties of the important metabolites. It must be also noted that the results indicated that the processing that takes place to produce milk powders influences the nutritional loss of peptides and lipids. HCA heatmaps showed that nutritional components were found to be in lower concentrations in reconstituted milk compared to ultra-high-temperature milk. Couvreur and Hurtaud [52] studied the parameters of fat globule characteristics (diameter, membrane surface, and yield), fat, protein, fatty acids, and calcium content in milk concerning diet composition, milking frequency, breed, stage of lactation, parity and residual/cisternal milk. Based on the principal components of PCA, HCA was performed which indicated 4 independent clusters of milk. A minor relationship was observed between fat content and fat globule diameter in milk, especially for the Normandy breed at the very end of the lactation. Dhankhar et al. [53] proposed a method to study the influence of season on the variability of sterols in different species' origins. Buffalo milk has a very different sterol profile compared to other animal species. In addition, seasonal variation affected especially cholesterol content compared to other minor sterols, and winter milk had a lower level of cholesterol compared to other seasons. The authors commented that the variation based on season was not able to be satisfactorily explained by PCA. However, HCA correctly grouped the 4 species of animals into 4 clusters by the sterol content. Squared Euclidean distance between objects was applied in HCA, to give the natural grouping of samples. The HCA dendrogram allowed the visualization of the similarity or dissimilarity of the measurements in 2D.

As can be observed, HCA is the main representative of the clustering methods. It is also important to note that after CA, most of the studies presented above proceeded to classification and/or regression analysis, which are presented in the next sections of this paper. Overall, in the aforementioned-studies, CA was used as a step to visualize the samples in clusters and to understand the interrelationships of the samples' datasets, before proceeding to supervised methods.

2.2. Classification

The capability to assign an object to a class on the basis of its characteristics belongs to the pattern recognition field. There are many methods to classify objects and one of the applications of chemometrics is the classification of objects in groups depending on their characteristics expressed as results of a set of measurements [54]. Classification methods could be distinguished into "discriminant" and "class-modeling" techniques (Table 2).

In the first case, the technique tries to discriminate among the object's groups dividing the model hyperspace into several regions equal to the number of classes and assigning each object to a specific region of the hyperspace on the base of its characteristics. In this way, each sample may belong to just one class. In the case of class modeling instead, the technique tries to model the analogies between objects of a class rather than observe the differences. So, each group of objects is modeled separately, and, at the end, an object could be assigned to one or more classes, or rejected as non-included in none of the classes (Figure 1).

Table 2. Main classification methods cited.

Classification Method	Extended Name	Abbreviation
Discriminant	Partial least squares-discriminant analysis	PLS-DA
	Orthogonal partial least squares-discriminant analysis	OPLS-DA
	One class-partial least squares	OC-PLS
	Quadratic discriminant analysis	QDA
	Random forest	RF
	Support vector machine	SVM
	Linear discriminant analysis	LDA
	k-nearest neighbors	kNN
	Extreme learning machine	ELM
	Ensemble of extreme learning machine	EELM
Class-modeling	Soft independent modeling of class analogy	SIMCA
	Data-driven soft independent modeling of class analogy	DD-SIMCA
	Unequal class models	UNEQ

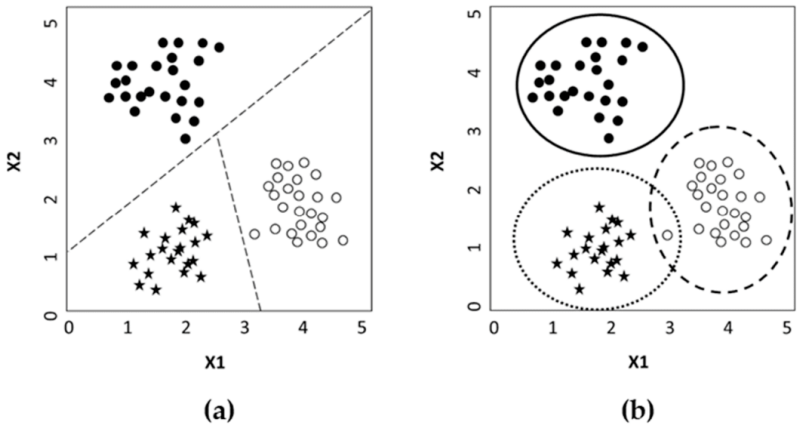


Figure 1. Example of difference between discriminant (**a**) modeling and (**b**) classification methods. In (**a**) the hyperspace is divided into regions equal to the category number.

In the discriminant classification, some methods may be counted: kNN, PLS-DA, LDA, and QDA. Instead, class-modeling techniques may be included: SIMCA, DD-SIMCA, and UNEQ [55].

Describing the details of all classification methods is out of the scope of this work, and here we will consider only the most used techniques (discriminant or class-modeling) applied to the milk and dairy product classification in milk adulteration in the last years.

A basic distinction between supervised and unsupervised classification techniques will be maintained. Supervised classification methods require some knowledge "a priori" of the classes and the method to assign or not assign samples to a certain class; in contrast, the unsupervised methods just classify samples on the base of their characteristics [56].

In recent years, the number of studies that use chemometrics to properly elaborate and interpret analytical results is largely increasing. The power of the chemometric technique is evident in all the cases where the output of an instrumental analytical technique is a spectrum, like in visible and/or infrared spectroscopy (VIS, VIS-NIRS, NIRS), nuclear magnetic resonance (NMR), or spectrometry (CG-MS, LC-MS).

Regarding classification used in milk adulteration, in the last five years, there have been several examples that used chemometrics and in Table 3 some relevant examples have been reported.

The use of chemometrics on instrumental data requires some preliminary steps, like data pre-processing or data dimension reduction. A short illustration of these steps has been reported below. In general, the application of a specific classification technique in place of another one depends on the data structure. In some cases, using one method rather than another one leads to the same results; in others, the application of a specific method could improve classification efficiency.

The classification statistical techniques most used in the last years for milk applications were PLS-DA as a pure classification technique and SIMCA as a class-modeling approach. Kamboj [57], for example, used PLS-DA to detect water adulteration in milk from NIRS spectra. Chung [58], working on isotope ratio data, used OPLS-DA to perform classification. The paper did not extensively explain the reason for this choice. Jin [59] used the least squares support vector machine (LS-SVM) for qualitative analysis of adulterated milk identification using 2D autocorrelation spectroscopic data. Karunathilaka [60] used Raman spectroscopy data from two different instruments and SIMCA for not-target classification to detect milk powder adulteration. Galvan [61], on data coming from low-cost spectroscopic devices (NIR and energy dispersive X-ray fluorescence—EDXRF), used more than one technique: PLS-DA for the EDXRF data and C-support vector classification (C-SVC) for NIR data. In the end, they concluded that DD-SIMCA was more useful to classify the samples with good accuracy (98.9%). Other two interesting uses of PLS-DA applied to NIR data were conducted by Ejeahalaka et al. [62] on cow's milk and by Di Donato et al. [63] on donkey's milk. DD-SIMCA is a one-class classification algorithm proposed in 2017 by Zontov [64]. The algorithm in the first phase is similar to the SIMCA algorithm, with a preliminary PCA. Then the PCA results were used to calculate the orthogonal distance and score distance for each object. These distances were then used to individuate a threshold limit value of the classification area. New samples were then classified in the orthogonal vs. score plot and assigned to the class when under the acceptance area defined for a given alpha value. Wang [65] evaluated four different classification methods (RF, LDA, SVM, and kNN) when dealing with milk authentication by infrared spectroscopy. To evaluate the best algorithm, the means of precision, accuracy, recall (true positive divided by the sum of true positive and false negative), and another parameter F1 (that together evaluate precision and recall) were calculated for each performance evaluation of all classes and for every classifier. The results indicate that RF had the best performance. In a work about image analysis [66] applied to recognize goat's milk (as a target class) from other milk species adulterants, two methods were tested: OC-PLS and DD-SIMCA. In this case, OC-PLS was not recommended and DD-SIMCA was preferred. Chen [67] used ELM and extreme ELM (EELM) to classify six types of milk of different brands analyzed by NIRS. ELM is a regression and classification algorithm. It is simple and efficient and extremely fast. Vargas [56] applied PLS on the voltammetric characterization of fresh cow's milk and from milk powder, using as Y the percentage of adulteration with reconstituted milk. Potocnik [68] in his paper used DA and OPLS-DA to elaborate data from isotopic ratios on types of milk to verify their geographical origins. Similarly, Xie [69] performed similar work on geographical discrimination of milk from Mongolia using isotope ratio, elements, and amino acids composition. In this paper, the chemometric analysis was performed with OPLS-DA. Tommasini [70], again using NMR, in this case, to classify the breed of cow, used PLS-DA analysis to distinguish between milk from different cow breeds, Friesian vs. autochthonous. PLS-DA and OPLS-DA, together with HCA and RF, were also cited by Sundelkide [71] to elaborate on the NMR spectra acquired in order to underline the importance and potentiality of the milk metabolomics studies. Segato et al. also used NMR to discriminate the metabolic profiles of different pasture-based alpine Asiago PDO cheeses [72]. To conclude the NMR overview, Yanibada [73] reported the application of OPLS-DA, preceded by an explorative PCA, to classify two groups of cows by NMR metabolomics. In Table 3 a synthesis of the more relevant papers identified has been reported.

To summarize, excluding PCA (mainly used to preliminarily study the problem), PLS-DA and OPLS-DA were the most used methods for classification in the recent papers on milk classification. The second most used have been SIMCA and DD-SIMCA, followed by many other various methods. The use of some classification techniques more than others could be attributed to different reasons: PLS-DA and OPLS-DA, the more used in the reviewed articles, are more known compared to some other more specific methods. The main reason for their popularity is probably linked to the fact that they are implemented in a lot of user-friendly commercial software, mainly used by non-expert users. It is advisable to use PLS-DA in place of LDA when the number of variables is higher than the number of samples and when the predictors are correlated. When classes are not balanced (i.e., the number of samples for each class is very different), better results are often obtained by class-modeling techniques, such as SIMCA. The choice of the proper classification method should also be influenced by their parametric or non-parametric nature: the former, such as LDA, assumes that the data follow a particular statistical distribution, so the model calculation becomes the calculation of the parameters of these distributions. The disadvantage of parametric techniques is that they can lead to big mistakes when starting assumptions fail to be verified. The advantage is that they make it easier to obtain the probability of obtaining a correct classification. On the other hand, non-parametric methods do not explicitly assume no statistical distribution (e.g., SIMCA, kNN, etc.).

Table 3. Recent studies (since 2018) involving classification methods related to milk adulteration.

Type of Milk	Target	Analytical Method(s)	Classification Method(s)	Reference
Cow	Classification	NIRS	EELM	Chen [67]
Cow	Organic milk geographical indication	Isotope ratio	OPLS-DA	Chung [58]
Cow	Authenticity	NMR	CDA	Segato [72]
Goat	Adulteration detection	Image analysis	OC-Classifier, OC-PLS, DD-SIMCA	dos Santos Pereira [66]
Cow	Quality	Chemical analysis, NIRS	PCA, SIMCA, PLS-DA	Ejeahalaka [62]
Various	Authenticity	NIRS, EDXRF	DD-SIMCA, PLS-DA, C-SVC	Galvan [61]
Cow	Adulteration	IR	LS-SVM	Jin [59]
Cow	Adulteration	NIRS	PCA, PLS	Kamboj [57]
Milk powder	Adulteration	Raman	PCA, SIMCA	Karunathilaka [60]
Cow	Geographical origin	Isotope ratio	ANOVA, DA, OPLS-DA, DD-SIMCA	Potočnik [68]
Cow	Authentication	Chemical analysis	PCA, OPLS-DA	Vargas [56]
Cow	Authentication	FTIR	PCA, kNN, SVM, RF, LDA	Wang [65]
Cow	Traceability	Chemical analysis, isotope ratio,	PCA, OPLS-DA	Xie [69]
Cow	Quality, breed classification	NMR	PLS, PLS-DA	Tomassini [70]

Table 3. *Cont.*

Type of Milk	Target	Analytical Method(s)	Classification Method(s)	Reference
Cow	Quality	NMR	PCA, PLS-DA, OPLS-DA, HCA, RF	Sundekilde [71]
Cow	Quality	NMR	PCA, OPLS-DA	Yanibada [73]
Donkey	Authentication	NIRS	PLS-DA, VSN, ASCA	Di Donato [63]

Abbreviations: ANOVA = analysis of variance, ASCA = ANOVA simultaneous component analysis, CDA = canonical discriminant analysis, C-SVC = C-classification support vector classifier, DA = discriminant analysis, DD-SIMCA = data-driven soft independent modeling of class analogy, EELM = ensemble of extreme learning machine, HCA = hierarchical cluster analysis, k-NN = k-nearest neighbors, LS-SVM = least squares support vector machine, LDA = linear discriminant analysis, OC = one-class classifier, OC-PLS = one-class partial least Squares, OPLS-DA = orthogonal partial least squares-discriminant analysis, PCA = principal component analysis, PLS = partial least squares, PLS-DA = partial least squares-discriminant analysis, RF = random forest, SVM = support vector machine, VSN = variable sorting for normalization.

2.3. Regression

Multivariate regression is widely used to quantify the concentration of adulterants in food matrices. In Table 4, the papers presented for this review in the last five years, with reference to regression methods, are listed.

The most popular multivariate regression method is certainly partial least squares (PLS) [74], as it is relatively simple to use and is implemented in a lot of statistical software, including instruments software (e.g., Opus). For this reason, in the last five years, PLS regression was used in more than three-quarters of the works on milk adulteration. The main advantage of PLS is its ability to handle data with many more variables than samples, specifically when these variables co-vary. The algorithm performs a simultaneous decomposition of both X (descriptors matrix) and Y (response matrix) matrices with the aim to maximize the covariance between the two matrices, computing at the same time latent variables (LVs) that explain the maximum variability of X. Due to its features, PLS is often used to treat spectral data, especially in the infrared region. In fact, with respect to other methods, such as chromatography, near- and mid-infrared spectroscopies (NIR and MIR, respectively) offer numerous practical advantages: they are fast, non-destructive, non-invasive, and relatively cheap techniques. Moreover, sample preparation is usually absent or extremely simple. The only drawback is the complex interpretation of the spectra, especially for NIR spectra, where differences in overtones and combination bands are difficult to detect and interpret. For this reason, the use of a simple multivariate tool for the extraction of relevant information is essential.

NIR spectroscopy is used to detect and quantify different kinds of adulterants: the most common and simple ones, such as water [57], urea [75–77], melamine [76–78], and sugar [79], and less common ones, such as sodium dodecyl sulfate (a milk surfactant) [80] or different vegetable oils added to yogurt [81]. Moreover, NIR spectroscopy is also used to detect specific adulterants for particular matrices as showed by Pandiselvam et al., where coconut milk residue was used to adulterate desiccated coconut powder [82], or by Di Donato et al., which used cow's milk as an adulterant in goat's milk samples [63].

MIR spectroscopy is also widely used coupled with PLS regression to detect and quantify adulterants in different milk samples. In several works, MIR was used to quantify the amount of cow's milk in more expensive milk types: buffalo [83,84], goat [85], and horse [86]. It was used to analyze coconut milk samples adulterated with water [87]. MIR spectrometers equipped with an ATR cell were employed to detect soya bean oil and common sugar [88], sucrose [89], and formalin [13] in cow's milk. The use of an ATR cell allows for minimizing sample preparation, as the penetration depth in the sample of IR radiation does not depend on sample thickness. Obviously, NIR and MIR spectra have to be properly pre-processed to minimize noise, scattering, and other undesirable contributions. Hence, it is good practice to build PLS models applying different combinations of pre-processing methods and compare the results to see which one provides the best prediction

performance. For instance, Temizkan et al. [81] tried different preprocessing options: normalization, smoothing, first derivative, second derivative, multiplicative scatter correction (MSC), and standard normal variate (SNV). These, together with the baseline correction, are the most common row pre-processing method used to treat NIR and MIR spectra.

Another spectroscopic technique coupled with PLS in the milk adulteration field is Raman spectroscopy, whose spectra rely on the light scattering of vibrating molecules. Raman spectroscopy was employed to find maltodextrin, sodium carbonate, and whey in bovine milk [90,91], as well as margarine, palm oil, and corn oil in cheeses made using adulterated milk samples [92,93].

Although in the majority of papers PLS regression is applied to vibrational spectroscopic data, in recent literature, there are also many applications with different techniques. Cyclic voltammetry, using a graphite/SiO_2 hybrid-working electrode, was employed to quantify reconstituted skim milk in cow's milk [11], electrochemical impedance spectroscopy was used to measure urea [36] whereas face fluorescence spectroscopy and laser-induced breakdown spectroscopy assessed the amount of bovine milk in buffalo milk [90] and ovine and caprine milk [94], respectively. Moreover, time-domain NMR [12] and opto-electronic nose [40] quantified formaldehyde in bovine milk. The versatility of this technique is one of the reasons why its presence is predominant among papers that deal with multivariate regression. Actually, in many papers, PLS is frequently compared with other two multivariate regression methods, i.e., multiple linear regression (MLR) [95] and principal component regression (PCR) [96]. Jaiswal et al. [85] and Gonçalves et al. [84] showed comparable results between PLS and MLR in quantifying adulterants with MIR spectroscopy. Conceição et al. [97] used MLR coupled with MIR spectroscopy to assess the amount of sodium bicarbonate, sodium hydroxide, hydrogen peroxide, starch, sucrose, and urea in cow's milk. However, the use of MLR is not recommended if the data matrix is ill-conditioned, namely has more variables (e.g., wavenumbers) than samples, and if those variables co-vary, as the regression model would be unstable. On the other hand, PCR is a more reliable method, since the variables are orthogonal (the ill-conditioned matrix problem has been overcome) and only relevant information in the original data matrix is considered, being based on PCA. Unlike PLS, in PCR the information in the response matrix (Y) is not taken into account when choosing the number of PCs. Moreover, for this reason, PLS has been habitually preferred to PCR. In some of the papers inspected for this review, these two methods were compared: on three occasions PLS provided the best prediction performances [13,86,89], whereas in one case the results obtained by the two methods were similar [87].

Throughout the years, the PLS algorithm has been modified by many authors to add features and make it more suitable for specific tasks (e.g., multiblock analysis, locally weighted models, etc.). One of the most famous extensions of PLS is orthogonal PLS (OPLS) [98], which removes the systematic variation from X that is not correlated (orthogonal) to Y. It was used by Delatour et al. [99] on data collected from eight different NIR and MIR miniature sensors to measure the amount of semicarbazide hydrochloride, ammonium sulfate, and cornstarch in skimmed milk powder [96]. Another different use of PLS regression, synergy interval PLS (siPLS) [100], has been used by Vinciguerra et al. to quantify cheese whey in cow's milk samples through MIR spectroscopy [38]. In this method, the MIR spectra were divided into different intervals (8, 16, 32, 64, and 128) with the same number of variables, applying a PLS on each interval. Furthermore, combinations of these intervals (two by two, three by three, and four by four) were also explored and PLS was performed for each combination. Hosseini et al. used the genetic algorithm PLS (GA-PLS) in order to perform an efficient variable selection before calculating the regression models [80]. Lastly, unfolded PLS with residual bilinearization (U-PLS/RBL) [101] coupled with fluorescence spectroscopy was used by Barreto et al. to quantify melamine in bovine milk [102]. Actually, U-PLS/RBL belongs to the family of multiway methods, similar to other techniques such as parallel factor analysis (PARAFAC) and multivariate curve resolution-alternating least squares (MCR-ALS), all based on obtaining pure profiles of the components present in

a mixture system. They are also called second-order calibration algorithms, as they can operate by decomposing the 3-way data matrix and then performing a regression between the resolved relative concentration of the constituents of interest and the corresponding reference concentration. Fluorescence spectroscopy provides excitation-emission matrices (EEMs) that can be resolved by those algorithms. According to de Araújo Gomes et al., U-PLS/RBL is particularly suitable to deal with fluorescence data, as it is able to model the inner filter effect that occurs in chemical fluorescence spectroscopy analysis systems [103]. Barreto et al. also used PARAFAC to quantify melamine, obtaining slightly better results than the ones achieved with U-PLS/RBL. PARAFAC [104] is a generalization of PCA to higher-order matrices, and its models furnish parameters (loadings) that describe the variability in the samples. Hence, MCR-ALS [105] was used by Zhao et al. on NIR data to compute calibration models for the simultaneous quantification of multiple adulterants (urea, melamine, and starch) [77]. In this case, MCR-ALS was used on classical 2-way data (i.e., NIR spectra), but the assumptions made earlier are valid. In general, MCR decomposes the data matrix into a bilinear model constraining the components' profiles to assure that the solution makes sense not only from a statistical point of view, but also chemically. ALS optimization explores the possible solutions through an iterative least square calculation until convergence is achieved.

Moving forward, some other less popular (but no worse) applications of multivariate regression techniques employed in the area of milk adulteration than PLS and its extensions can be found in the literature. Artificial neural network (ANN) regression methods, namely generalized regression-NN [106] and back propagation-ANN [107], were used to assess the amount of melamine, wheat flour, and corn flour in milk powder samples [108] and acidity in cow's milk samples [109], respectively, both through Raman spectroscopy. Least squares support vector machine (LS-SVM) [110] was applied on both NIR and dielectric spectroscopic data to quantify mature bovine milk in colostrum samples [111] and on MIR data to assess cheese whey in bovine milk [38], providing better results than PLS. A generalized linear model with lasso regularization (GLM-Lasso) [112] coupled with MALDI-TOF mass spectroscopy provides better results than PLS too, in this case, to detect bovine milk in caprine and ovine milk [113]. Ehsani et al. applied boosted regression tree (BRT) [114] on NIR spectra collected by a portable spectrometer for a fast water quantification in cow's milk [115]. The presence of water in cow's milk was also inspected by Asefa et al. [116], who proposed a procedure based on digital image analysis coupled with extreme gradient boosting (XGBoost) [117].

To sum up, the most-used technique for multivariate regression in the field of milk adulteration is by far PLS, as it is relatively simple to use and is present in much commercial software. In most cases, proper use of PLS regression is enough to obtain good prediction performances, but in the case of a more complex data structure, it is worth trying more advanced techniques. The use of the many extensions of PLS can be useful to increase the signal-to-noise ratio, to compute prediction models only with the most relevant variables, or to deal with 3-way data. More expert users sometimes use other kinds of multivariate regression methods, such as ANN or SVM. In some cases, they provide slightly better results than PLS, but in many other cases, the results are comparable.

Table 4. Recent studies (2018–2022) involving regression methods related to milk adulteration.

Type of Milk	Target	Analytical Method(s)	Regression Method(s)	Reference
Cow milk	Water	NIR	PLS	[57]
Cow milk	Urea	NIR	PLS	[75]
Fat-filled milk powder	Melamine, urea	NIR	PLS	[76]
Goat milk powder	Melamine, urea, starch	NIR	PLS, MCR-ALS	[77]
Milk powder—infant formula	Melamine, vanillin	NIR HSI	PLS	[78]
Cow milk	Sugar	NIR	PLS	[79]
Cow milk	Anionic surfactant (SDS)	NIR, MIR (ATR)	PLS, GA-PLS	[80]
Yogurt	Margarine, sunflower oil, corn oil, hydrogenated vegetable oil	NIR, MIR	PLS	[81]
Desiccated coconut powder	Coconut milk	Vis-NIR	PLS	[82]
Donkey milk	Cow milk	NIR	PLS	[73]
Buffalo milk	Cow milk	MIR	PLS	[83]
Buffalo milk	Cow milk	MIR	PLS, MLR	[84]
Goat milk	Cow milk	MIR, Raman	PLS	[85]
Horse milk	Cow milk, goat milk	MIR	PLS, PCR	[86]
Coconut milk	Water	MIR	PLS, PCR	[87]
Cow milk	Soya bean oil, sugar	MIR (ATR)	PLS, MLR	[88]
Cow milk	Sucrose	MIR (ATR)	PLS, PCR	[89]
Cow milk	Formalin	MIR (ATR)	PLS, PCR	[13]
Cow milk	Maltodextrin, sodium carbonate, whey	Raman	PLS	[90]
Cow milk	Whey	Raman	PLS	[91]
White ultra-filtered cheese	Margarine, palm oil, and corn oil	Raman	PLS	[92]
Cow milk	Reconstituted skim milk powder	Cyclic voltammetry	PLS	[11]
Cow milk	Urea	Electrochemical impedance spectroscopy	PLS	[36]
Buffalo milk	Cow milk	Face fluorescence spectroscopy	PLS	[93]
Ovine and caprine milk	Cow milk	Laser-induced breakdown spectroscopy	PLS	[94]
Cow milk	Formaldehyde	TD-NMR	PLS	[12]
Cow milk	Formaldehyde	Opto-electronic nose	PLS	[40]
Cow milk	Sodium bicarbonate, sodium hydroxide, hydrogen peroxide, starch, sucrose, urea	MIR (ATR)	MLR	[97]
Skimmed milk powder	Semicarbazide hydrochloride, ammonium sulfate, cornstarch	NIR (miniature spectral devices)	OPLS	[99]

Table 4. *Cont.*

Type of Milk	Target	Analytical Method(s)	Regression Method(s)	Reference
Cow milk	Whey	MIR	PLS, siPLS, LS-SVM	[38]
Cow milk	Melamine	Fluorescence spectroscopy	PARAFAC, U-PLS/RBL	[102]
Milk powder	Melamine, wheat flour, corn flour	Raman	GRNN	[108]
Cow milk	Acidity	Raman	PLS, BP-ANN	[109]
Colostrum	Mature cow milk	NIR, dielectric spectroscopy	PLS, LS-SVM	[111]
Ovine milk and caprine milk	Cow milk	MALDI-TOF-MS	PLS, GLM-Lasso	[113]
Cow milk	Water	NIR (portable)	BRT	[115]
Cow milk	Water	Digital image analysis	XGBoost	[116]

Abbreviations: ATR = attenuated total reflection, BP-ANN = back propagation artificial neural networks, BRT = boosted regression trees, GA-PLS = genetic-algorithm partial least squares, GLM-Lasso = generalized linear model with lasso regularization, GR-NN = generalized regression neural networks, HSI = hyperspectral imaging, LS-SVM = least squares support vector machine, MALDI-TOF-MS = matrix-assisted laser desorption ionization time-of-flight mass spectrometry, MCR-ALS = multivariate curve resolution alternating least squares, MIR = mid-infrared, MLR = multiple linear regression, NIR = near-infrared, OPLS = orthogonal partial least squares, PARAFAC = parallel factor analysis, PCR = principal component regression, PLS = partial least squares, siPLS = synergy interval partial least squares, TD-NMR = time-domain nuclear magnetic resonance, U-PLS/RBL = unfolded partial least squares with residual bilinearization, Vis = visible, XGBoost = extreme gradient boosting.

3. Steps for Development and Validation of a Chemometric Approach

It is difficult to define a precise pipeline for the correct development and validation of a chemometric approach for authentication purposes. This chapter tries to face the fundamental steps, covering the sampling procedure, considering the analytical source of data, the model calibration and validation, and the main figure of merits useful for model evaluation (Figure 2).

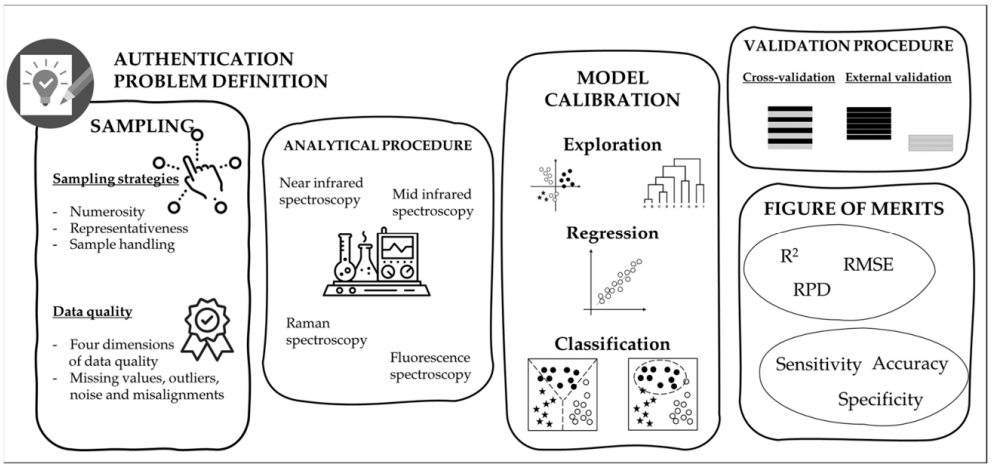


Figure 2. Schematic representation of the main steps useful to develop and validate a chemometric approach.

3.1. Correct Sampling Procedure

3.1.1. Sampling Strategies

No matter the chemometric model to be performed, according to the developed strategy goal, it is mandatory to perform a proper sampling strategy. Behind the word "proper" there are a set of extremely challenging standpoints that should consider the nature of the sample, the statistical representativeness, the analytical chemistry principles, and the quality and the management of the obtained datasets. Sampling procedures are very important to assure the robustness and reliability of the developed chemometric models. However, no well-defined sampling protocols exist so far for fingerprint techniques.

When addressing the nature of the sample, a relevant emphasis should be placed on the heterogeneity of a food matrix, together with the wide possibility of frauds, from the adulteration, i.e., the presence of an undeclared substance to the geographical origin discrimination, passing through the substitution of ingredients or commodities. In any case, the source of the samples, i.e., the provider, must be extremely reliable when addressing an authentication issue. They must be of provable provenance to assure they are authentic or not; thus, it would be advisable to obtain them from the producer rather than buying at retail markets [21].

For instance, the collection of commercial samples from local grocery stores to study goat's milk adulteration by cow's milk [85] could be inappropriate. Indeed, the commercial milk already passed to technological operation (heat treatments, fat separation, homogenization); thus, it would be more representative of real fraud to mix the different types of milk before any unit operation. This is what was done by Spina et al. [83], who described in detail the farmers, the breeds, and the sampling period and batches. Furthermore, they strengthened their experimental plan by planning a randomized pairing of cow and buffalo milk to obtain 17 adulteration levels.

Pandiselvam et al. [82] also adopted the strategy of ad hoc sample preparation. They prepared different adulterated samples by adding coconut milk residue to desiccated coconut powder. Even though the sample numerosity was quite high, i.e., 20 samples prepared for ten adulteration levels (from 0 to 100% *w/w*), it seems that the raw materials used to prepare the standard samples were always the same, thus not covering all the possible sources of variability. The variability of simulated adulterated samples was better covered by de Oliveira Mendes et al. [88], who considered six samples of milk from different producers to be adulterated with sweet whey prepared at a laboratory scale at eight adulteration levels.

From a statistical standpoint, the size and the representativeness of the sample collection must be considered [32] to obtain samples spanning all the sources of variability associated with the application of the model [118]. Different strategies described by the theory of sampling (ToS) could be followed to guarantee representative sampling and appropriate analytical quality [119]. A power analysis could be performed to establish the adequate number of samples required and to reduce the technical and biological variability. When a wide variability should be covered in a limited set of measures, design of experiments (DoE) techniques could be applied to obtain statistically valid data; the advantages of these approaches are well described by Peris-Díaz & Krężel [120].

In the literature there are examples of poor sampling strategies; for example, there are works considering a number of samples that is too low to be representative from both a technological/chemical and statistical point of view [12,38,87].

From an analytical point of view, the sample handling in terms of conservation prior to analysis, preparation, and analytical replicates should be faced to circumscribe the intrinsic variability. This is quite a challenging issue which has been clearly pointed out by Kemsley, et al. [121], and too often poorly described in the revised literature.

Finally, to sum up the useful sampling strategy to be adopted, the approach proposed by the "Five Ws" iterative interrogative technique could be winning. The first W to be clearly set is the goal of the developed approach, i.e., why, and the definition of the authentication issue to be addressed. Then, it is appropriate to cover the personnel and

instrument variability (who), together with the definition of sample unit, the number of samples, handling procedure, representativeness, balanced/not balanced datasets, and possible development/availability of trusted samples (what). Moreover, the range of time (when)—which could refer to seasons, harvesting years, vintages, product aging, and so on—should be adequately covered. Finally, the investigation of the effects of the area of origin and/or the processing steps (where) should be faced.

3.1.2. Data Quality

The quality of the collected raw data strongly influences the data processing and the model quality. This is highly dependent on the instrumentation characteristics and related analytical methodology. The review by Szymanska [122] deeply described the four main dimensions of data quality (accuracy, completeness, timeliness, and consistency) and their characteristics. The most common artifacts generated by quality collection failures are missing values, outliers, noise, and misalignments. According to the type, there are strategies for their detection and deletion, substitution, or correction [122]. However, in most of the literature, little attention is given to the description of these strategies, which are hopefully applied to assess and monitor the quality of the collected data before the chemometric model construction.

3.2. Pre-Processing

An exception is the description of data pre-processing, which is generally reported as a winning strategy to remove irrelevant sources of variation, such as instrumental and experimental artifacts due to the employed analytical method. However, there are still authors who miss the preprocessing description, such as Kamboj et al. (2020) [57], or just mention an automatic strategy applied by the software. Different preprocessing strategies are available; in-depth information is given by Engel et al. [123]. Every specific dataset has specific features; thus, the definition of a rule of thumb to define which preprocessing strategy is more appropriate is impossible.

In any case, the spectroscopic data requires a pre-processing step before the statistical data analysis to remove or minimize variability in the spectra not related to the sample's characteristics. It will be clear that pre-processing cannot generate information, but only help to extract proper information already existing in the data. Moreover, incorrect use of pre-processing may cause a loss of information. Pre-treatment should be well calibrated to minimize the effects of "noise" such as optical phenomena, effects of temperature changes, light scattering, baseline shift or trends, and so on.

Most of the revised works, especially the ones dealing with infrared data, apply different preprocessing strategies, such as smoothing, standard normal variate (SNV) or multiplicative scatter correction (MSC), and derivatives alone or in combination [87,99,124]. Later on, they select the most appropriate one to solve the specific adulteration issue based on the performance criteria obtained in the developed models. However, it is important not to apply all of them by default without looking back at their effect on the data. Indeed, it should be considered that an inappropriate transformation can cause alterations to data quality, driving relevant consequences on model outcomes. A must-read tutorial concerning pre-processing has been written by Oliveri et al. [125].

Between the papers explored, some different approaches have been found in NIR pre-processing. Ejeahalaka [62] performed a comparison between two different approaches: first, no pre-processing at all, and second, extended multiplicative signal correction (EMSC) on a selected part of the spectrum. In Galvan [61] some different pre-processing methods were tested before a mean centering for all: (1) raw data, (2) Savitzky–Golay smoothing (third-order polynomial and 21 window points), (3) standard normal variate (SNV), (4) multiplicative scatter correction (MSC), (5) first and second derivative with Savitzky–Golay smoothing, (6) SNV plus first and second derivative, and (7) MSC plus first and second derivative. At the end, the best performance (evaluated by RMSE of the calculated models) was obtained by the application of the first derivative with smoothing (pre-processing 5).

Wang [65] used three pre-processing steps: (1) mean centering, (2) first, and (3) second-order Savitzky–Golay derivative, selecting at the end the first-order derivative as the better pre-processing method.

Kamboj [57] did not indicate which pre-processing was used. Not mentioning the pre-processing step should be avoided because this step implies some assumptions on the nature of the data set variability, and it is crucial that these assumptions are well understood and appropriate. An innovative approach was reported by Di Donato [63] in a study on donkey milk. NIR data were used to identify and quantify cow adulteration in expensive donkey milk. In this case, the pre-processing was done by variable sorting for normalization (VSN), a recent scatter correction technique [126] that estimates the weight of wavelengths that are or are not related to scattering effects instead of that related to the response of interest. Not-related wavelengths were not considered in the successive step. In this way, it is possible to obtain an improvement in signal and model interpretation.

Karunathilaka [60] in an application of Raman spectroscopy cites different spectral pre-processing to remove fluorescence and laser fluctuations, including Savitzky–Golay first and second derivatives and standard normal variate (SNV), choosing at the end the second derivative.

3.3. Data Reduction

The analysis of spectroscopic results is a typical example in which the dimension of the analytical part of the dataset (n columns) is much higher than the number of samples (m rows), normally thousands of columns vs. tens or hundreds of rows. So, to avoid elaboration problems and to select just the variables relevant for the statistical analysis, a variable selection step is often evaluated. Reviewing in detail all the possible algorithms is out of scope, considering their relevant number; thus, here we only report the ones used in the evaluated papers. Between them, just a few used a data reduction algorithm. For example, Chen [67] on NIRS data used an extension of the ReliefF filter algorithm [74]. ReliefF filter works on multiple classes, building a weight vector that indicates for each feature (wavelengths in the NIRS case) how important it is to explain the differences between samples of different classes. Wang [65] instead used just an observation of the first two PCA loadings as the criterion to understand relevant wavelengths, but it was unclear if just the relevant wavelengths in the subsequent classification step were used.

3.4. Use of Robust Validation Procedures

Before detailing the possible validation procedures, it is essential to consider the quality of the calibration. Taking for granted that the data representativeness and numerosity must be guaranteed according to the defined purpose, it is relevant not to overfit or underfit the model calibration.

Model validation is frequently addressed by iterative validation procedures, such as cross-validation. In the considered papers, the most used cross-validation strategy is leave-one-out, to whatever degree it should be avoided for its over-optimistic results, especially in the case of exhaustive sampling procedures [13,75,83]. Indeed, it means that during the iterative recalculation of the model just one sample at a time is removed; this way the robustness of the model is poorly investigated. None of the work internally validates models with other iterative procedures such as Monte Carlo, Jackknife, holdout, or bootstrapping.

The use of internal validation is often justified when a low number of samples is at disposal. In these situations, it can be unaffordable to exclude 30–40% of the collected data to be used as a test set. Westad and Marini [127] suggest this strategy when the number of samples is smaller than 40.

Moreover, the internal validation procedures are fundamental insights to study the model stability, identify the main sources of variation, and improve model performance, i.e., by setting model dimensionality [128]. This was the approach followed by Ejeahalaka et al. [76] for both SIMCA and PLS model development. It is important to notice

that the correct model dimensionality is fundamental for predicting the test set; if the model dimensionality is incorrect, the performance criteria/figure of merit may not be a good estimate of future samples, as reported by Westad and Marini [127]. For instance, the results obtained from internal validation give insights about model overfitting due to the selection of a huge number of components/variables, which means fitting too much of the data so that also the measurement noise is interpreted as a relevant effect.

Then, it is the time to use robust, mandatory validation procedures in order to guarantee reliable and reproducible results. Usually, the available samples are divided into two subsets: a training (or calibration) set to be used for building the model, and a test set used to evaluate its validity [20] in terms of quality and generalization ability [129]. The division should guarantee that the calibration set covers the whole variability domain to obtain reliable results. The dataset split could be performed arbitrarily—according to the acquired knowledge of the data, randomly, or designed by sampling strategies—such as the Kennard and Stone algorithm, Duplex, D-optimality criterion, and K-means or Kohonen mapping; for more details about the differences among the strategies and their effects refer to Westad and Marini [127].

Infrequently, the experimental structure is considered for data splitting. This was the case for Genis et al. [92] who considered 15 concentrations of fat in the calibration set, and 11 concentrations of fat as validation data set when developing methods for the identification of foreign lipid types and adulteration ratio in milk. Most of the revised papers apply random sample selection to build the test set considering from 40 to 20% of the whole data. Among the designed sampling strategies, the Kennard and Stone algorithm is the one mostly used. However, in many cases no information is provided for dataset splitting, thus making the model robustness evaluation difficult.

In any case, it would be advisable to use a fully independent set of data to test the model; for example, considering a different production batch, a different time of the year, or a different harvesting year.

This option will represent the ideal procedure for model validation, anyway it should be set to guarantee the samples' diversity if possible, or at least their mutual independence [130].

If someone argues it is still not enough, we can reply as suggested by Westad and Marini [127]: "Another way to overcome the problem of using the same criterion to select a subset of variables and the error (i.e., cross-validation) is to divide the objects into a calibration, a validation and a verification set, where the verification set is the 'proof of the pudding'".

Each step of model development (i.e., calibration, cross-validation, and external validation) should be properly evaluated by diagnostic metrics (i.e., Figures of Merits), which are discussed in the next session.

3.5. Performance Criteria/Figure of Merits

Before mentioning the performance criteria useful for regression evaluation, it is important to have enough information to evaluate the quality of the collected data. In particular specific information must be reported about the numerosity of the data, their variability (i.e., mean, median, and standard deviation), the nature of the measure (instrumentation used), the removed outliers (and adopted strategy), the regression algorithm employed (mainly PLS, OPLS, PCR, MLR, LSSVM, SWM, ANN, GLM-Lasso, and so on), or the classification approach (mainly PLS-DA and OPLS-DA for the pure classification, and SIMCA and DD-SIMCA for the class-modeling techniques), the characteristics of the model development steps (calibration, internal- and external validation), the potential data pre-treatments, and the selected components/latent variables [131]. Last but not least, the information about the reference method employed to determine the specific compound and the associated error, i.e., the standard error of the laboratory (SEL), or the standard error of the test (SET), must be reported [131]. Having a clear idea of the variance covered by the data and the error of the reference analysis would be crucial to judge the results obtained by the regression model obtained. Indeed, the accuracy of chemometric model

predictors depends on the repeatability of the reference methods and it combines both the error of the reference measure and the error of the fingerprint analysis [132].

3.5.1. R^2 (Coefficient of Determination) and RMSE (Root Mean Squared Error)

The main effective tests used to evaluate multivariate regression models are R^2, SEP, and the RPD. R^2, the coefficient of determination, is commonly used to evaluate regression models in every development step. It is quite relevant to compare the different coefficients of determination obtained in calibration, cross-validation, and prediction to understand the model stability. It would be better to evaluate the R^2 adjusted, which corrects for the number of explanatory terms in relation to the number of data points.

The coefficient of determination (R^2) is, in its most general definition, computed by:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (1)$$

where SS_{res} is the sum of squares of residuals for measurements y_i and mean of observed data ($\bar{Y}$) and SS_{tot} is the total sum of squares.

The R^2 adjusted is:

$$R^2_{adj} = 1 - \frac{n-1}{1-k-1}\frac{SS_{res}}{SS_{tot}} \quad (2)$$

where n is the number of observations and k is the number of independent variables.

However, the evaluation of R^2 alone is not exhaustive: there may be models with high coefficient values, thus describing high data variability, but with high error, expressed as root mean square error. To determine the reasonability of RMSE value it should be compared to measurement errors such as reference method, reproducibility error, historical data, and so on.

The RMSE is computed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} \quad (3)$$

where n is the number of observations, y_i is the predicted value and $\hat{y}$ is the actual value.

If divided by the standard deviation of the experimental values it is obtained the normalized RMSE (nRMSE), which is an unbiased measurement for model predictions.

Good error estimation was performed for the models developed by Genis [92]. They calculated the relative error of standard deviation (RSD) and relative error of prediction (REP) together with the limit of detection (LOD) and the limit of quantification (LOQ) in the regression model intended for fat authenticity in milk for ultra-filtered white cheese.

The use of both criteria, R^2 and RMSE, is relevant especially in cases of high range of variability of the considered compound; in this case, it could be plausible to obtain a model with higher R^2, but accompanied by higher RMSE, if compared with a dataset with limited range of variability. Generally speaking, "wide" calibration could be less precise, but more dangerous is a too-narrow calibration which will be valid just for the case understudy [132].

The ratio between the SD and the RMSE is referred to as ratio percentage deviation (RPD). It can be seen as a performance criterion like R^2, even if RPD is a ratio of SD, whereas R^2 is a ratio of variance. Its calculation is present in few papers dealing with milk adulteration [13,82,84,89,91], but its use can give an immediate insight to evaluate the predictions as well as to compare models predicting different compounds [132]. There are different papers that give an interpretation of model performance according to RPD values, among them the one of Williams [133] which defines six levels of performance. In the considered works the RPD was always quite high. Indeed, very good prediction capabilities were reached by the MLR model for buffalo's milk authenticity verification developed by Gonçalves et al. [84]; the RPD was 7.9. When developing a PLS regression on the same data it improved to 9.0, thus demonstrating the excellent performance of mid-infrared spectroscopy to assess cow's milk levels in buffalo's milk. The model developed by

Pandiselvam et al. [82] for the detection of adulteration with coconut powder also achieved excellent performance, resulting in an RPD of 11. Excellent performances were found by Balan et al. [13] when developing a PLS model to predict formalin in cow's milk, reaching an RPD above 8. Also, the RPD of the PLS models developed by Balan et al. [89] was high (13.4), demonstrating an excellent prediction capability of sucrose in milk, thus being able to detect sucrose addition intended to increase total solid content as well as the sweet taste. Similarly, de Oliveira Mendes et al. [91] developed a PLS model for whey quantification in raw milk by Raman spectroscopy obtaining an RPD of 13.9.

In any case, where RPD is not reported as a model parameter, it can be calculated directly from the R^2 such as $1/-(1 - R^2)$.

Bellon-Maurel et al. [134] proposed to substitute RPD with a new index, RPIQ (ratio of performance to IQ). The index is based on quartiles, thus better representing the population distribution. They found out that, in sample sets with skewed distribution, the RPD is not a good approach for SEP standardization according to population spread, whereas the RPIQ index, in which standard deviation is replaced by IQ (=Q3 − Q1), better considers the spread of the population. However, none of the works considered here applied this figure of merit.

3.5.2. Specificity and Sensitivity, and Graphical Representations

The performance of classification models is assessed by verifying if samples belonging to the class of interest are designated as true positives (TP) or false negatives (FN), as well as if samples not belonging to the class of interest are labeled as false positives (FP) or true negatives (TN) [20]. Just to recall the theory, TP defines the samples recognized to belong to the class a priori assigned, FN are samples erroneously rejected, FP are samples erroneously assigned to the class, and TN are samples correctly refused.

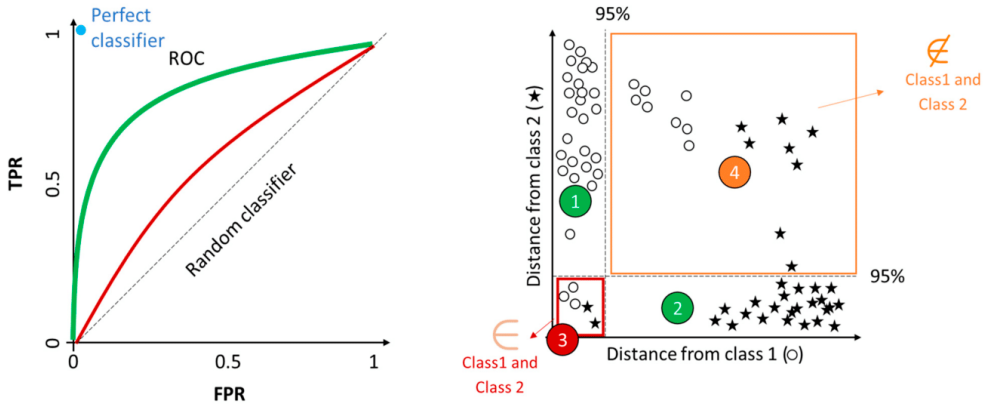
From their assignments, it is possible to calculate the sensitivity and sensibility of the method. Sensitivity is the true positive rate (TPR), computed as TP/(TP + FN). Specificity is the true negative rate (TNR), computed as TN/(TN + FP).

The graphical tool used to represent the performance criteria of a discriminant model is the receiver operating characteristic (ROC) curves (Figure 3a). The plot represents a two-axis Cartesian space, with the horizontal axis reporting FPR, and the vertical axis the TPR. The dashed diagonal represents the performance of a random classifier. Two examples of classifiers (green and red) are shown, representing good and scarce results, respectively. The curves are built by connecting with a line the experimental outcomes. This tool is useful to compare the performances of models obtained with different parameter settings, such as the threshold value. A detailed analysis of ROC curves is discussed by Oliveri [20].

If discriminant methods can be applied only to solve multi-class situations, class modeling can be used to address both multi-class and one-class problems.

When performing a class-modeling analysis it could be useful to evaluate the results with a graphical representation, so Coomans' plots (Figure 3b). In a two-class problem, the two axes represent the distances of samples from the models of Class 1 (○) and Class 2 (star), respectively. The two dashed lines correspond to the critical acceptance levels for each model at the defined confidence level (normally 95%). Samples of the two classes are projected as scatter points, with coordinates indicating the relative similarity with the two models in the four sectors defined in the plot. In sector 1 it is possible to find samples accepted only by Class 1 (○); in sector 2 it is possible to find samples accepted only by Class 2 (star). Both sectors include samples defined as TP for the a priori defined class.

In sector 3 are positioned samples accepted by both models; indeed, since models for each class are independently built, class spaces may overlap. Lastly, in sector 4 it is possible to observe samples rejected by both models, which highlights that the used variables do not completely resolve the class space. They prevent the forced (but possibly wrong) classification of samples that may occur in discriminant approaches [20].



(**a**) (**b**)

Figure 3. Graphical tools to represent classification model performance. (**a**) Receiver operating characteristic (ROC) curves; (**b**) Coomans' plot.

4. Methods for Rapid and On-Site Detection to Combat Milk Adulteration

The dairy industry as well as regulatory bodies are looking for simple and rapid methods for the detection of milk adulteration [135]. Lateral flow immunoassays (LFIAs) have been used as in situ screening tools to monitor food raw material quality as they provide rapid results [136]. LFIAs have been developed, among other applications [137], for the detection and quantification of mycotoxins [138], such as aflatoxin M1 [139]. LFIAs have been also used for the detection of adulteration of milk with melamine [140]. In a very recent study adulteration of cow's milk with buffalo's milk was detected by an on-site carbon nanoparticle-based lateral flow immunoassay in 10 min, with the sensitivity of the test being 5%, i.e., 5% adulteration of cow's milk with buffalo's milk, proving that this tool is suitable for rapid detection of adulteration [135].

Another novel technology for the rapid detection of milk adulteration is DNAFoil. It is a portable, fully self-administered, on-site DNA test that does not require the use of expensive PCR equipment or laboratory setups to confirm the detection of milk adulteration within a short period of time. The efficiency of the DNAFoil kit used to detect the vegetable material in milk products (DNAFoil UniPlant) was confirmed using real-time PCR assays. The results showed that using 24 µL of DNAFoil UniPlant master mix, a 17.5 min reaction time allowed the detection of 10% adulteration of liquid cow's milk by wheat flour [141].

Moreover, an electronic nose (e-nose) system is being evolved for the falsification detection of milk and dairy products in a reliable and rapid way [142]. This technology avoids the disadvantages of chromatography, spectrometry, and chemical methods with high costs and long cycle times [143]. Adulteration of bovine milk with formaldehyde, based on aldehydes and ketones, was examined by electronic nose by Mostafapour et al. [40]. In another investigation, the identification of trace amounts of detergent powder in raw milk using a customized low-cost electronic nose was achieved [144].

5. Conclusions

An overview of the different chemometric techniques (from clustering to classification and regression applied to several analytical data) has been presented along with spectroscopy, chromatography, and electrochemical sensors as well as rapid and on-site detection devices in the fight against milk adulteration and fraud. HCA is the main representative of the clustering methods. The classification of objects in groups depending on their characteristics expressed as results of a set of measurements is one of the applications of chemometrics. Classification methods were distinguished into "discriminant" and

"class-modeling" techniques. The classification statistical techniques mostly employed in the last few years for milk applications were PLS-DA as a pure classification technique and SIMCA as a class-modeling approach. Multivariate regression is widely used to quantify the concentration of adulterants in food matrices and was deeply described.

Finally, the steps which should be followed to develop a chemometric model to face adulteration issues were carefully presented with the required critical discussion describing sampling strategies, pre-processing, data reduction, and use of robust validation procedures along with performance criteria/figure of merits.

All chemometric methods, supervised and unsupervised, had fundamental results in order to serve the goals of each research study. It cannot be concluded which chemometric method is the best, as each dataset is unique and different. Robustness is usually more related to supervised methods, but unsupervised methods are also important in the field. Usually, the availability and access to each chemometric method are the variables that influence their specific selection. With regard to the field of milk adulteration, it is clear that, in most cases, the simplest methods are enough to obtain good results. However, even the simplest methods are in some cases used improperly, making the results obtained inconsistent.

Author Contributions: Conceptualization, S.G., M.T., A.D., S.A., T.V. and L.S.; methodology, S.G., M.T., A.D., S.A., T.V. and L.S.; writing—original draft preparation, S.G., M.T., A.D., S.A., T.V. and L.S.; writing—review and editing, S.G., M.T., A.D., S.A., T.V. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Foods for the financial support provided.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Uncu, A.O.; Uncu, A.T. A barcode-DNA analysis method for the identification of plant oil adulteration in milk and dairy products. *Food Chem.* **2020**, *326*, 126986. [CrossRef] [PubMed]
2. Zhang, T.; Wu, X.; Wu, B.; Dai, C.; Fu, H. Rapid authentication of the geographical origin of milk using portable near-infrared spectrometer and fuzzy uncorrelated discriminant transformation. *J. Food Process. Eng.* **2022**, *45*, e14040. [CrossRef]
3. Ye, H.; Yang, J.; Xiao, G.; Zhao, Y.; Li, Z.; Bai, W.; Zeng, X.; Dong, H. A comprehensive overview of emerging techniques and chemometrics for authenticity and traceability of animal-derived food. *Food Chem.* **2023**, *402*, 134216. [CrossRef] [PubMed]
4. Nascimento, C.F.; Santos, P.M.; Pereira-Filho, E.R.; Rocha, F.R. Recent advances on determination of milk adulterants. *Food Chem.* **2017**, *221*, 1232–1244. [CrossRef] [PubMed]
5. Moore, J.C.; Spink, J.; Lipp, M. Development and application of a database of food ingredient fraud and economically motivated adulteration from 1980 to 2010. *J. Food Sci.* **2012**, *77*, R118–R126. [CrossRef]
6. Giglioti, R.; Polli, H.; Azevedo, B.T.; Katiki, L.M.; Vercesi Filho, A.E. Detection and quantification of adulteration in milk and dairy products: A novel and sensitive qPCR-based method. *Food Chem. Mol. Sci.* **2022**, *4*, 100074. [CrossRef]
7. Teixeira, J.L.D.P.; Carames, E.T.D.S.; Baptista, D.P.; Gigante, M.L.; Pallone, J.A.L. Vibrational spectroscopy and chemometrics tools for authenticity and improvement the safety control in goat milk. *Food Control* **2020**, *112*, 107105. [CrossRef]
8. Du, L.; Lu, W.; Cai, Z.J.; Bao, L.; Hartmann, C.; Gao, B.; Yu, L.L. Rapid detection of milk adulteration using intact protein flow injection mass spectrometric fingerprints combined with chemometrics. *Food Chem.* **2018**, *240*, 573–578. [CrossRef]
9. Motta, T.C.; Hoff, R.B.; Barreto, F.; Andrade, R.B.S.; Lorenzini, D.M.; Meneghini, L.Z.; Pizzolato, T.M. Detection and confirmation of milk adulteration with cheese whey using proteomic-like sample preparation and liquid chromatography–electrospray–tandem mass spectrometry analysis. *Talanta* **2014**, *120*, 498–505. [CrossRef]
10. Qin, C.; Liu, L.; Wang, Y.; Leng, T.; Zhu, M.; Gan, B.; Xie, J.; Yu, Q.; Chen, Y. Advancement of omics techniques for chemical profile analysis and authentication of milk. *Trends Food Sci. Technol.* **2022**, *127*, 114–128. [CrossRef]
11. Nikolaou, P.; Deskoulidis, E.; Topoglidis, E.; Kakoulidou, A.T.; Tsopelas, F. Application of chemometrics for detection and modeling of adulteration of fresh cow milk with reconstituted skim milk powder using voltammetric fingerpriting on a graphite/SiO_2 hybrid electrode. *Talanta* **2020**, *206*, 120223. [CrossRef] [PubMed]
12. Coimbra, P.T.; Bathazar, C.F.; Guimarães, J.T.; Coutinho, N.M.; Pimentel, T.C.; Neto, R.P.C.; Esmerino, E.A.; Freitas, M.Q.; Silva, M.C.; Tavares, M.I.B.; et al. Detection of formaldehyde in raw milk by time domain nuclear magnetic resonance and chemometrics. *Food Control* **2020**, *110*, 107006. [CrossRef]
13. Balan, B.; Dhaulaniya, A.S.; Jamwal, R.; Sodhi, K.K.; Kelly, S.; Cannavan, A.; Singh, D.K. Application of Attenuated Total Reflectance-Fourier Transform Infrared (ATR-FTIR) spectroscopy coupled with chemometrics for detection and quantification of formalin in cow milk. *Vib. Spectrosc.* **2020**, *107*, 103033. [CrossRef]

14. Wasnik, P.G.; Menon, R.R.; Sivaram, M.; Nath, B.S.; Balasubramanyam, B.V.; Manjunatha, M. Development of mathematical model for prediction of adulteration levels of cow ghee with vegetable fat using image analysis. *J. Food Sci. Technol.* **2019**, *56*, 2320–2325. [CrossRef]
15. Roy, M.; Doddappa, M.; Yadav, B.K.; Jaganmohan, R.; Sinija, V.R.; Manickam, L.; Sarvanan, S. Detection of soybean oil adulteration in cow ghee (clarified milk fat): An ultrafast study using flash gas chromatography electronic nose coupled with multivariate chemometrics. *J. Sci. Food Agric.* **2022**, *102*, 4097–4108. [CrossRef]
16. Vatavali, K.; Kosma, I.; Louppis, A.; Gatzias, I.; Badeka, A.V.; Kontominas, M.G. Characterisation and differentiation of geographical origin of Graviera cheeses produced in Greece based on physico-chemical, chromatographic and spectroscopic analyses, in combination with chemometrics. *Int. Dairy J.* **2020**, *110*, 104799. [CrossRef]
17. Aleixandre-Tudo, J.L.; Castello-Cogollos, L.; Aleixandre, J.L.; Aleixandre-Benavent, R. Chemometrics in food science and technology: A bibliometric study. *Chemom. Intell. Lab. Syst.* **2022**, *222*, 104514. [CrossRef]
18. Kamal, M.; Karoui, R. Analytical methods coupled with chemometric tools for determining the authenticity and detecting the adulteration of dairy products: A review. *Trends Food Sci. Technol.* **2015**, *46*, 27–48. [CrossRef]
19. Gómez-Caravaca, A.M.; Maggio, R.M.; Cerretani, L. Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil. A review. *Anal. Chim. Acta* **2016**, *913*, 1–21. [CrossRef]
20. Oliveri, P. Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues—A tutorial. *Anal. Chim. Acta* **2017**, *982*, 9–19. [CrossRef]
21. McGrath, T.F.; Haughey, S.A.; Patterson, J.; Fauhl-Hassek, C.; Donarski, J.; Alewijn, M.; van Ruth, S.; Elliott, C.T. What are the scientific challenges in moving from targeted to non-targeted methods for food fraud testing and how can they be addressed? Spectroscopy case study. *Trends Food Sci. Technol.* **2018**, *76*, 38–55. [CrossRef]
22. Cubero-Leon, E.; Penalver, R.; Maquet, A. Review on metabolomics for food authentication. *Food Res. Int.* **2014**, *60*, 95–10711.
23. Hanganu, A.; Chira, N. When detection of dairy food fraud fails: An alternate approach through proton nuclear magnetic resonance spectroscopy. *J. Dairy Sci.* **2021**, *104*, 8454–8466. [CrossRef] [PubMed]
24. Souhassou, S.; Bassbasi, M.; Hirri, A.; Kzaiber, F.; Oussama, A. Detection of camel milk adulteration using Fourier transformed infrared spectroscopy FT-IR coupled with chemometrics methods. *Int. Food Res. J.* **2018**, *25*, 1213–1218.
25. Wang, X.; Esquerre, C.; Downey, G.; Henihan, L.; O'Callaghan, D.; O'Donnell, C. Feasibility of discriminating dried dairy ingredients and preheat treatments using mid-infrared and Raman Spectroscopy. *Food Anal. Methods* **2018**, *11*, 1380–1389. [CrossRef]
26. Karunathilaka, S.R.; Yakes, B.J.; He, K.; Chung, J.K.; Mossoba, M. Non-targeted NIR spectroscopy and SIMCA classification for commercial milk powder authentication: A study using eleven potential adulterants. *Heliyon* **2018**, *4*, e00806. [CrossRef]
27. Da Silva Dias, L.; da Silva Junior, J.C.; Felício, A.L.D.S.M.; de França, J.A. A NIR photometer prototype with integrating sphere for the detection of added water in raw milk. *IEEE Trans. Instrum. Meas.* **2018**, *67*, 2812–2819. [CrossRef]
28. Windarsih, A.; Rohman, A.; Irnawati; Riyanto, S. The Combination of Vibrational Spectroscopy and Chemometrics for Analysis of Milk Products Adulteration. *Int. J. Food Sci.* **2021**, *2021*, 8853358. [CrossRef]
29. de Lima, A.B.S.; Batista, A.S.; de Jesus, J.C.; de Jesus Silva, J.; de Araújo, A.C.M.; Santos, L.S. Fast quantitative detection of black pepper and cumin adulterations by near-infrared spectroscopy and multivariate modeling. *Food Control* **2020**, *107*, 106802. [CrossRef]
30. Jiménez-Carvelo, A.M.; González-Casado, A.; Bagur-González, M.G.; Cuadros-Rodríguez, L. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity—A review. *Food Res. Int.* **2019**, *122*, 25–39. [CrossRef]
31. Teixeira, J.L.d.P.; Caramês, E.T.d.S.; Baptista, D.P.; Gigante, M.L.; Pallone, J.A.L. Rapid adulteration detection of yogurt and cheese made from goat milk by vibrational spectroscopy and chemometric tools. *J. Food Compost. Anal.* **2021**, *96*, 103712. [CrossRef]
32. Ramirez-Lopez, L.; Schmidt, K.; Behrens, T.; Van Wesemael, B.; Demattê, J.A.; Scholten, T. Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma* **2014**, *226*, 140–150. [CrossRef]
33. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv. CSUR* **1999**, *31*, 264–323. [CrossRef]
34. Bratchell, N. Chapter 6 Cluster Analysis. *Data Handl. Sci. Technol.* **1992**, *9*, 179–208.
35. Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O.P.; Tiwari, A.; Er, M.J.; Ding, W.; Lin, C. A review of clustering techniques and developments. *Neurocomputing* **2017**, *267*, 664–681. [CrossRef]
36. Minetto, T.A.; França, B.D.; da Silva Dariz, G.; Veiga, E.A.; Galvão, A.C.; da Silva Robazza, W. Identifying adulteration of raw bovine milk with urea through electrochemical impedance spectroscopy coupled with chemometric techniques. *Food Chem.* **2022**, *385*, 132678. [CrossRef]
37. Cirak, O.; Icyer, N.C.; Durak, M.Z. Rapid detection of adulteration of milks from different species using Fourier Transform Infrared Spectroscopy (FTIR). *J. Dairy Res.* **2018**, *85*, 222–225. [CrossRef]
38. Vinciguerra, L.L.; Marcelo, M.C.; Motta, T.; Meneghini, L.Z.; Bergold, A.M.; Ferrão, M.F. Chemometric tools and FTIR-ATR spectroscopy applied in milk adulterated with cheese whey. *Química Nova* **2019**, *42*, 249–254. [CrossRef]
39. Ezhilan, M.; Gumpu, M.B.; Ramachandra, B.L.; Nesakumar, N.; Babu, K.J.; Krishnan, U.M.; Rayappan, J.B.B. Design and development of electrochemical biosensor for the simultaneous detection of melamine and urea in adulterated milk samples. *Sens. Actuators B Chem.* **2017**, *238*, 1283–1292. [CrossRef]

40. Mostafapour, S.; Gharaghani, F.M.; Hemmateenejad, B. Converting electronic nose into opto-electronic nose by mixing MoS_2 quantum dots with organic reagents: Application to recognition of aldehydes and ketones and determination of formaldehyde in milk. *Anal. Chim. Acta* **2021**, *1170*, 338654. [CrossRef]
41. Li, Q.; Yu, Z.; Zhu, D.; Meng, X.; Pang, X.; Liu, Y.; Frew, R.; Chen, H.; Chen, G. The application of NMR-based milk metabolite analysis in milk authenticity identification. *J. Sci. Food Agric.* **2017**, *97*, 2875–2882. [CrossRef] [PubMed]
42. Sowmya, N.; Ponnusamy, V. Development of spectroscopic sensor system for an IoT application of adulteration identification on milk using machine learning. *IEEE Access* **2021**, *9*, 53979–53995. [CrossRef]
43. Souza, S.O.; Santos, V.S.; Santos, E.S.; Ávila, D.V.L.; Nascimento, C.C.; Costa, S.S.L.; Garcia, C.A.B.; Araujo, R.G.O. Evaluation of the mineral content in milk and yogurt types using chemometric tools. *Microchem. J.* **2018**, *143*, 1–8. [CrossRef]
44. Rodríguez-Bermúdez, R.; López-Alonso, M.; Miranda, M.; Fouz, R.; Orjales, I.; Herrero-Latorre, C. Chemometric authentication of the organic status of milk on the basis of trace element content. *Food Chem.* **2018**, *240*, 686–693. [CrossRef]
45. Zain, S.M.; Behkami, S.; Bakirdere, S.; Koki, I.B. Milk authentication and discrimination via metal content clustering–A case of comparing milk from Malaysia and selected countries of the world. *Food Control* **2016**, *66*, 306–314. [CrossRef]
46. Xu, S.; Zhao, C.; Deng, X.; Zhang, R.; Qu, L.; Wang, M.; Ren, S.; Wu, H.; Yue, Z.; Niu, B. Determining the geographical origin of milk by multivariate analysis based on stable isotope ratios, elements and fatty acids. *Anal. Methods* **2021**, *13*, 2537–2548. [CrossRef]
47. Amenzou, N.; Hamid, M.; Fouad, T.; Elyahyaoui, A.; Elghali, T.; Elmoqrani, L.; Mahmoud, E. *Stable Isotope Ratios in Dairy Products (Milk) as New Tool to Determine Their Different Origins in Morocco*; Joint FAO/IAEA Centre of Nuclear Techniques in Food and Agriculture, Food Safety and Control Section: Vienna, Austria, 2022; pp. 60–69, 128, ISBN 978-92-0-124822-0. ISSN 1011-4289. CONTRACT MOR 18051.
48. Podkolzin, I.; Solovev, A. *Application of Stable Isotope Techniques and Elemental Analysis to Confirm Geographical Origin of Milk Produced in the Russian Federation*; IAEA: Vienna, Austria, 2022.
49. Karrar, E.; Mohamed Ahmed, I.A.; Huppertz, T.; Wei, W.; Jin, J.; Wang, X. Fatty acid composition and stereospecificity and sterol composition of milk fat from different species. *Int. Dairy J.* **2022**, *128*, 105313. [CrossRef]
50. Bhumireddy, S.R.; Rocchetti, G.; Pallerla, P.; Lucini, L.; Sripadi, P. A combined targeted/untargeted screening based on GC/MS to detect low-molecular-weight compounds in different milk samples of different species and as affected by processing. *Int. Dairy J.* **2021**, *118*, 105045. [CrossRef]
51. Tan, D.; Zhang, X.; Su, M.; Jia, M.; Zhu, D.; Kebede, B.; Wu, H.; Chen, G. Establishing an untargeted-to-MRM liquid chromatography–mass spectrometry method for discriminating reconstituted milk from ultra-high temperature milk. *Food Chem.* **2021**, *337*, 127946. [CrossRef]
52. Couvreur, S.; Hurtaud, C. Relationships between milks differentiated on native milk fat globule characteristics and fat, protein and calcium compositions. *Animal* **2017**, *11*, 507–518. [CrossRef]
53. Dhankhar, J.; Sharma, R.; Indumathi, K. A comparative study of sterols in milk fat of different Indian dairy animals based on chemometric analysis. *J. Food Meas. Charact.* **2020**, *14*, 2538–2548. [CrossRef]
54. Marini, F. Classification methods in chemometrics. In Proceedings of the Mediterraneum Meeting, Ventotene, Italy, 1–4 June 2008.
55. Derde, M.P.; Massart, D.L. UNEQ: A disjoint modelling technique for pattern recognition based on normal distribution. *Anal. Chim. Acta* **1986**, *184*, 33–51. [CrossRef]
56. Vargas-Bello-Pérez, E.; Gomez-Cortes, P.; Geldsetzer-Mendoza, C.; Sol Morales, M.; Toro-Mujica, P.; Fellenberg, M.A.; Ibanez, R.A. Authentication of retail cheeses based on fatty acid composition and multivariate data analysis. *Int. Dairy J.* **2018**, *85*, 280–284. [CrossRef]
57. Kamboj, U.; Kaushal, N.; Mishra, S.; Munjal, N. Application of Selective Near Infrared Spectroscopy for Qualitative and Quantitative Prediction of Water Adulteration in Milk. *Mater. Today Proc.* **2020**, *24*, 2449–2456. [CrossRef]
58. Chung, I.M.; Kim, J.K.; Yang, Y.J.; An, Y.J.; Kim, S.Y.; Kwon, C.; Kim, S.H. A case study for geographical indication of organic milk in Korea using stable isotope ratios-based chemometric analysis. *Food Control* **2020**, *107*, 106755. [CrossRef]
59. Jin, H.; Dong, G.M.; Wu, H.Y.; Yang, Y.R.; Huang, M.Y.; Wang, M.Y.; Yang, R.J. Identification of adulterated milk based on auto-correlation spectra. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2023**, *286*, 121987. [CrossRef]
60. Karunathilaka, S.R.; Yakes, B.J.; He, K.; Brückner, L.; Mossoba, M.M. First use of handheld Raman spectroscopic devices and on-board chemometric analysis for the detection of milk powder adulteration. *Food Control* **2018**, *92*, 137–146. [CrossRef]
61. Galvan, D.; Lelis, C.A.; Effting, L.; Melquiades, F.L.; Bona, E.; Conte-Junior, C.A. Low-cost spectroscopic devices with multivariate analysis applied to milk authenticity. *Microchem. J.* **2022**, *181*, 107746. [CrossRef]
62. Ejeahalaka, K.K.; On, S.L.W. Chemometric studies of the effects of milk fat replacement with different proportions of vegetable oils in the formulation of fat-filled milk powders: Implications for quality assurance. *Food Chem.* **2019**, *295*, 198–205. [CrossRef]
63. 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. *J. Food Compos. Anal.* 2022, *in press*. [CrossRef]
64. Zontov, Y.V.; Rodionova, O.Y.; Kucheryavskiy, S.V.; Pomerantsev, A.L. DD-SIMCA—A MATLAB GUI tool for data driven SIMCA approach. *Chemom. Intell. Lab. Syst.* **2017**, *167*, 23–28. [CrossRef]
65. Wang, Y.T.; Ren, H.B.; Liang, W.Y.; Jin, X.; Yuan, Q.; Liu, Z.R.; Chen, D.M.; Zhang, Y.H. A novel approach to temperature-dependent thermal processing authentication for milk by infrared spectroscopy coupled with machine learning. *J. Food Eng.* **2021**, *311*, 110740. [CrossRef]

66. Dos Santos Pereira, E.V.; de Sousa Fernandes, D.D.; de Almeida, L.F.; Sucupira Maciel, M.I.; Gonçalves Dias Diniz, P.H. Goat milk authentication by one-class classification of digital image-based fingerprint signatures: Detection of adulteration with cow milk. *Microchem. J.* **2022**, *180*, 107640. [CrossRef]
67. Chen, H.; Tan, C.; Lin, Z.; Wua, T. Classification of different liquid milk by near-infrared spectroscopy and ensemble modelling. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2021**, *251*, 119460. [CrossRef]
68. Potočnik, D.; Nečemer, M.; Perišić, I.; Jagodic, M.; Mazej, D.; Camin, F.; Eftimov, T.; Strojnik, L.; Ogrinc, N. Geographical verification of Slovenian milk using stable isotope ratio, multielement and multivariate modelling approaches. *Food Chem.* **2020**, *326*, 126958. [CrossRef]
69. Xie, L.; Zhao, S.; Rogers, K.M.; Xia, Y.; Zhang, B.; Suo, R.; Zhao, Y. A case of milk traceability in small-scale districts-Inner Mongolia of China by nutritional and geographical parameters. *Food Chem.* **2020**, *316*, 126332. [CrossRef]
70. Tommasini, A.; Curone, G.; Solè, M.; Capuani, G.; Sciubba, F.; Conta, G.; Miccheli, G.; Vigo, D. NMR-based metabolomics to evaluate the milk composition from Friesian and autochtonous cows of Northern Italy at different lactation times. *Nat. Prod. Res.* **2019**, *33*, 1085–1091. [CrossRef]
71. Sundekilde, U.K.; Larsen, L.B.; Bertram, C. NMR-Based milk metabolomics. *Metabolites* **2013**, *3*, 204–222. [CrossRef]
72. Segato, S.; Caligiani, A.; Contiero, B.; Galaverna, G.; Bisutti, V.; Cozzi, G. 1H NMR metabolic profile to discriminate pasture based alpine Asiago PDO cheeses. *Animals* **2019**, *9*, 722. [CrossRef]
73. Yanibada, B.; Boudra, H.; Debrauwer, L.; Martin, C.; Morgavi, D.P.; Canlet, C. Evaluation of sample preparation methods for NMR-based metabolomics of cow milk. *Heliyon* **2018**, *4*, e00856. [CrossRef]
74. Wold, S.; Martens, H.; Wold, H. The multivariate calibration problem in chemistry solved by PLS method. In *Matrix Pencils*; Springer: Berlin/Heidelberg, Germany, 1983; pp. 286–293.
75. Mabood, F.; Ali, L.; Boque, R.; Abbas, G.; Jabeen, F.; Haq, Q.M.I.; Hussain, J.; Hamaed, A.M.; Naureen, Z.; Al-Nabhani, M.; et al. Robust Fourier transformed infrared spectroscopy coupled with multivariate methods for detection and quantification of urea adulteration in fresh milk samples. *Food Sci. Nutr.* **2020**, *8*, 5249–5258. [CrossRef] [PubMed]
76. Ejeahalaka, K.K.; On, S.L. Effective detection and quantification of chemical adulterants in model fat-filled milk powders using NIRS and hierarchical modelling strategies. *Food Chem.* **2020**, *309*, 125785. [CrossRef] [PubMed]
77. Zhao, X.; Wang, Y.; Liu, X.; Jiang, H.; Zhao, Z.; Niu, X.; Li, C.; Pang, B.; Li, Y. Single-and Multiple-Adulterants Determinations of Goat Milk Powder by NIR Spectroscopy Combined with Chemometric Algorithms. *Agriculture* **2022**, *12*, 434. [CrossRef]
78. Zhao, X.; Li, C.; Zhao, Z.; Wu, G.; Xia, L.; Jiang, H.; Wang, T.; Chu, X.; Liu, J. Generic models for rapid detection of vanillin and melamine adulterated in infant formulas from diverse brands based on near-infrared hyperspectral imaging. *Infrared Phys. Technol.* **2021**, *116*, 103745. [CrossRef]
79. Kamboj, U.; Kaushal, N.; Jabeen, S. Near Infrared Spectroscopy as an efficient tool for the Qualitative and Quantitative Determination of Sugar Adulteration in Milk. *J. Phys. Conf. Ser.* **2020**, *1531*, 12024. [CrossRef]
80. Hosseini, E.; Ghasemi, J.B.; Daraei, B.; Asadi, G.; Adib, N. Application of genetic algorithm and multivariate methods for the detection and measurement of milk-surfactant adulteration by attenuated total reflection and near-infrared spectroscopy. *J. Sci. Food Agric.* **2021**, *101*, 2696–2703. [CrossRef]
81. Temizkan, R.; Can, A.; Dogan, M.A.; Mortas, M.; Ayvaz, H. Rapid detection of milk fat adulteration in yoghurts using near and mid-infrared spectroscopy. *Int. Dairy J.* **2020**, *110*, 104795. [CrossRef]
82. Pandiselvam, R.; Mahanti, N.K.; Manikantan, M.R.; Kothakota, A.; Chakraborty, S.K.; Ramesh, S.V.; Beegum, P.S. Rapid detection of adulteration in desiccated coconut powder: Vis-NIR spectroscopy and chemometric approach. *Food Control* **2022**, *133*, 108588. [CrossRef]
83. Spina, A.A.; Ceniti, C.; Piras, C.; Tilocca, B.; Britti, D.; Morittu, V.M. Mid-Infrared (MIR) Spectroscopy for the quantitative detection of cow's milk in buffalo milk. *J. Anim. Sci. Technol.* **2022**, *64*, 531–538. [CrossRef]
84. Gonçalves, B.H.R.; Silva, G.J.; Jesus, J.C.D.; Conceição, D.G.; Santos, L.S.; Ferrão, S.P. Fast verification of buffalo's milk authenticity by mid-infrared spectroscopy, analytical measurements and multivariate calibration. *J. Braz. Chem. Soc.* **2020**, *31*, 1453–1460. [CrossRef]
85. Yaman, H. A rapid method for detection adulteration in goat milk by using vibrational spectroscopy in combination with chemometric methods. *J. Food Sci. Technol.* **2020**, *57*, 3091–3098. [CrossRef] [PubMed]
86. Arifah, M.F.; Nisa, K.; Windarsih, A.; Rohman, A. The Application of FTIR Spectroscopy and Chemometrics for the Authentication Analysis of Horse Milk. *Int. J. Food Sci.* **2022**, *2022*, 7643959. [CrossRef] [PubMed]
87. Sitorus, A.; Muslih, M.; Cebro, I.S.; Bulan, R. Dataset of adulteration with water in coconut milk using FTIR spectroscopy. *Data Br.* **2021**, *36*, 107058. [CrossRef]
88. Jaiswal, P.; Jha, S.N.; Kaur, J.; Hg, R. Rapid detection and quantification of soya bean oil and common sugar in bovine milk using attenuated total reflectance–fourier transform infrared spectroscopy. *Int. J. Dairy Technol.* **2018**, *71*, 292–300. [CrossRef]
89. Balan, B.; Dhaulaniya, A.S.; Jamwal, R.; Yadav, A.; Kelly, S.; Cannavan, A.; Singh, D.K. Rapid detection and quantification of sucrose adulteration in cow milk using Attenuated total reflectance-Fourier transform infrared spectroscopy coupled with multivariate analysis. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *240*, 118628. [CrossRef] [PubMed]
90. Tian, H.; Chen, S.; Li, D.; Lou, X.; Chen, C.; Yu, H. Simultaneous detection for adulterations of maltodextrin, sodium carbonate, and whey in raw milk using Raman spectroscopy and chemometrics. *J. Dairy Sci.* **2022**, *105*, 7242–7252. [CrossRef] [PubMed]

91. De Oliveira Mendes, T.; Rodrigues, B.V.M.; Porto, B.L.S.; da Rocha, R.A.; de Oliveira, M.A.L.; de Castro, F.K.; de Carvalho dos Anjos, V.; Bell, M.J.V. Raman Spectroscopy as a fast tool for whey quantification in raw milk. *Vib. Spectrosc.* **2020**, *111*, 103150. [CrossRef]
92. Genis, D.O.; Sezer, B.; Durna, S.; Boyaci, I.H. Determination of milk fat authenticity in ultra-filtered white cheese by using Raman spectroscopy with multivariate data analysis. *Food Chem.* **2021**, *336*, 127699. [CrossRef]
93. Ullah, R.; Khan, S.; Ali, H.; Bilal, M. Potentiality of using front face fluorescence spectroscopy for quantitative analysis of cow milk adulteration in buffalo milk. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *225*, 117518. [CrossRef]
94. Sezer, B.; Durna, S.; Bilge, G.; Berkkan, A.; Yetisemiyen, A.; Boyaci, I.H. Identification of milk fraud using laser-induced breakdown spectroscopy (LIBS). *Int. Dairy J.* **2018**, *81*, 1–7. [CrossRef]
95. Lai, T.L.; Robbins, H.; Wei, C.Z. Strong consistency of least squares estimates in multiple regression II. *J. Multivar. Anal.* **1979**, *9*, 343–361. [CrossRef]
96. Mandel, J. Use of the singular value decomposition in regression analysis. *Am. Stat.* **1982**, *36*, 15–24.
97. Conceição, D.G.; Gonçalves, B.H.R.; Hora, F.F.D.; Faleiro, A.S.; Santos, L.S.; Ferrão, S.P. Use of FTIR-ATR spectroscopy combined with multivariate analysis as a screening tool to identify adulterants in raw milk. *J. Braz. Chem. Soc.* **2019**, *30*, 780–785. [CrossRef]
98. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom. J. Chemom. Soc.* **2002**, *16*, 119–128. [CrossRef]
99. Delatour, T.; Becker, F.; Krause, J.; Romero, R.; Gruna, R.; Längle, T.; Panchaud, A. Handheld spectral sensing devices should not mislead consumers as far as non-authentic food is concerned: A case study with adulteration of milk powder. *Foods* **2021**, *11*, 75. [CrossRef]
100. Nørgaard, L.; Hahn, M.T.; Knudsen, L.B.; Farhat, I.A.; Engelsen, S.B. Multivariate near-infrared and Raman spectroscopic quantifications of the crystallinity of lactose in whey permeate powder. *Int. Dairy J.* **2005**, *15*, 1261–1270. [CrossRef]
101. Öhman, J.; Geladi, P.; Wold, S. Residual bilinearization. Part 1: Theory and algorithms. *J. Chemom.* **1990**, *4*, 79–90. [CrossRef]
102. Barreto, M.C.; Braga, R.G.; Lemos, S.G.; Fragoso, W.D. Determination of melamine in milk by fluorescence spectroscopy and second-order calibration. *Food Chem.* **2021**, *364*, 130407. [CrossRef] [PubMed]
103. De Araújo Gomes, A.; Schenone, A.V.; Goicoechea, H.C.; de Araújo, M.C.U. Unfolded partial least squares/residual bilinearization combined with the Successive Projections Algorithm for interval selection: Enhanced excitation-emission fluorescence data modeling in the presence of the inner filter effect. *Anal. Bioanal. Chem.* **2015**, *407*, 5649–5659. [CrossRef]
104. Bro, R. PARAFAC. Tutorial and applications. *Chemom. Intel. Lab. Syst.* **1997**, *38*, 149–171. [CrossRef]
105. De Juan, A.; Tauler, R. Multivariate curve resolution (MCR) from 2000: Progress in concepts and applications. *Crit. Rev. Anal. Chem.* **2006**, *36*, 163–176. [CrossRef]
106. Kelwade, J.P.; Salankar, S.S. Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series. In Proceedings of the 2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI), Kolkata, India, 8–10 January 2016; IEEE: New York, NY, USA, 2016; pp. 454–458.
107. Yabunaka, K.I.; Hosomi, M.; Murakami, A. Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. *Water Sci. Technol.* **1997**, *36*, 89–97. [CrossRef]
108. Wang, Y.; Liu, Q. Fast identification of powdered milk adulteration by generalized regression neural network algorithm. In *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2021)*; SPIE: Bellingham, WA, USA, 2022; Volume 12168, pp. 717–724.
109. Huang, W.; Fan, D.; Li, W.; Meng, Y.; Liu, T.C.Y. Rapid evaluation of milk acidity and identification of milk adulteration by Raman spectroscopy combined with chemometrics analysis. *Vib. Spectrosc.* **2022**, *123*, 103440. [CrossRef]
110. Suykens, J.A.K.; Vandewalle, J. Least square support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
111. Yang, K.; An, C.; Zhu, J.; Guo, W.; Lu, C.; Zhu, X. Comparison of near-infrared and dielectric spectra for quantitative identification of bovine colostrum adulterated with mature milk. *J. Dairy Sci.* **2022**, *105*, 8638–8649. [CrossRef] [PubMed]
112. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
113. Rysova, L.; Cejnar, P.; Hanus, O.; Legarova, V.; Havlik, J.; Nejeschlebova, H.; Nemeckova, I.; Jedelska, R.; Bozik, M. Use of MALDI-TOF MS technology to evaluate adulteration of small ruminant milk with raw bovine milk. *J. Dairy Sci.* **2022**, *105*, 4882–4894. [CrossRef]
114. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
115. Ehsani, S.; Dastgerdy, E.M.; Yazdanpanah, H.; Parastar, H. Ensemble classification and regression techniques combined with portable near infrared spectroscopy for facile and rapid detection of water adulteration in bovine raw milk. *J. Chemom.* 2022, *Early View*.
116. Asefa, B.G.; Hagos, L.; Kore, T.; Emire, S.A. Feasibility of Image Analysis Coupled with Machine Learning for Detection and Quantification of Extraneous Water in Milk. *Food Anal. Methods* **2022**, *15*, 3092–3103. [CrossRef]
117. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
118. Schlesier, K.; Fauhl-Hassek, C.; Forina, M.; Cotea, V.; Kocsi, E.; Schoula, R.; van Jaarsveld, F.; Wittkowaki, R. Characterisation and determination of the geographical origin of wines. Part I: Overview. *Eur. Food Res. Technol.* **2009**, *230*, 13. [CrossRef]
119. Esbensen, K.H.; Paoletti, C.; Thiex, N. Representative sampling for food and feed materials: A critical need for food/feed safety. *J. AOAC Int.* **2015**, *98*, 249–251. [CrossRef] [PubMed]

120. Peris-Díaz, M.D.; Krężel, A. A guide to good practice in chemometric methods for vibrational spectroscopy, electrochemistry, and hyphenated mass spectrometry. *Trends Anal. Chem.* **2021**, *135*, 116157. [CrossRef]
121. Kemsley, E.K.; Defernez, M.; Marini, F. Multivariate statistics: Considerations and confidences in food authenticity problems. *Food Control* **2019**, *105*, 102–112. [CrossRef]
122. Szymańska, E. Modern data science for analytical chemical data—A comprehensive review. *Anal. Chim. Acta* **2018**, *1028*, 1–10. [CrossRef] [PubMed]
123. Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J.J.; Downey, G.; Blanchet, L.; Buydens, L.M. Breaking with trends in pre-processing? *Trends Anal. Chem.* **2013**, *50*, 96–106. [CrossRef]
124. Yang, Y.; Hettinga, K.A.; Erasmus, S.W.; Pustjens, A.M.; van Ruth, S.M. Opportunities for fraudsters: When would profitable milk adulterations go unnoticed by common, standardized FTIR measurements? *Food Res. Int.* **2020**, *136*, 109543. [CrossRef]
125. Oliveri, P.; Malegori, C.; Simonetti, R.; Casale, M. The impact of signal pre-processing on the final interpretation of analytical outcomes—A tutorial. *Anal. Chim. Acta* **2019**, *1058*, 9–17. [CrossRef] [PubMed]
126. Rabatel, G.; Marini, F.; Walczak, B.; Roger, J.M. VSN: Variable sorting for normalization. *J. Chemom.* **2020**, *34*, e3164. [CrossRef]
127. Westad, F.; Marini, F. Validation of chemometric models–a tutorial. *Anal. Chim. Acta* **2015**, *893*, 14–24. [CrossRef]
128. Rajamanickam, V.; Babel, H.; Montano-Herrera, L.; Ehsani, A.; Stiefel, F.; Haider, S.; Presser, B.; Knapp, B. About model validation in bioprocessing. *Processes* **2021**, *9*, 961. [CrossRef]
129. Smilde, A.; Bro, R.; Geladi, P. *Multi-Way Analysis in Chemistry and Related Fields*; John Wiley & Sons, Ltd.: New York, NY, USA, 2004; Volume 240, pp. 260–280.
130. Daszykowski, M.; Walczak, B.; Massart, D.L. Representative subset selection. *Anal. Chim. Acta* **2022**, *468*, 91–103. [CrossRef]
131. Williams, P.; Dardenne, P.; Flinn, P. Tutorial: Items to be included in a report on a near infrared spectroscopy project. *J. Near Infrared Spectrosc.* **2017**, *25*, 85–90. [CrossRef]
132. Bittante, G.; Patel, N.; Cecchinato, A.; Berzaghi, P. Invited review: A comprehensive review of visible and near-infrared spectroscopy for predicting the chemical composition of cheese. *J. Dairy Sci.* **2022**, *105*, 1817–1836. [CrossRef]
133. Williams, P.C. Implementation of near-infrared technology. In *Near-Infrared Technology in the Agricultural and Food Industries*; American Association of Cereal Chemist Press: St. Paul, Minnesota, USA, 2001; pp. 145–169.
134. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.M.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends Anal. Chem.* **2010**, *29*, 1073–1081. [CrossRef]
135. Sharma, R.; Verma, A.; Shinde, N.; Mann, B.; Gandhi, K.; Wichers, J.H.; van Amerongen, A. Adulteration of cow's milk with buffalo's milk detected by an on-site carbon nanoparticles-based lateral flow immunoassay. *Food Chem.* **2021**, *351*, 129311. [CrossRef] [PubMed]
136. Agriopoulou, S.; Stamatelopoulou, E.; Varzakas, T. Advances in analysis and detection of major mycotoxins in foods. *Foods* **2020**, *9*, 518. [CrossRef]
137. Di Nardo, F.; Chiarello, M.; Cavalera, S.; Baggiani, C.; Anfossi, L. Ten Years of Lateral Flow Immunoassay Technique Applications: Trends, Challenges and Future Perspectives. *Sensors* **2021**, *21*, 5185. [CrossRef]
138. Agriopoulou, S.; Stamatelopoulou, E.; Varzakas, T. Advances in Occurrence, Importance, and Mycotoxin Control Strategies: Prevention and Detoxification in Foods. *Foods* **2020**, *9*, 137. [CrossRef]
139. Wang, C.; Peng, J.; Liu, D.-F.; Xing, K.-Y.; Zhang, G.-G.; Huang, Z.; Cheng, S.; Zhu, F.F.; Duan, M.L.; Zhang, K.Y.; et al. Lateral flow immunoassay integrated with competitive and sandwich models for the detection of aflatoxin M1 and *Escherichia coli* O157:H7 in milk. *J. Dairy Sci.* **2018**, *101*, 8767–8777. [CrossRef]
140. Yue, X.; Pan, Q.; Zhou, J.; Ren, H.; Peng, C.; Wang, Z.; Zhang, Y. A simplified fluorescent lateral flow assay for melamine based on aggregation induced emission of gold nanoclusters. *Food Chem.* **2022**, *385*, 132670. [CrossRef]
141. El Sheikha, A.F. DNAFoil: Novel technology for the rapid detection of food adulteration. *Trends Food Sci. Technol.* **2019**, *86*, 544–552. [CrossRef]
142. Roy, M.; Yadav, B.K. Electronic nose for detection of food adulteration: A review. *J. Food Sci. Technol.* **2022**, *59*, 846–858. [CrossRef] [PubMed]
143. Tian, H.; Chen, B.; Lou, X.; Yu, H.; Yuan, H.; Huang, J.; Chen, C. Rapid detection of acid neutralizers adulteration in raw milk using FGC E-nose and chemometrics. *J. Food Meas. Charact.* **2022**, *16*, 2978–2988. [CrossRef]
144. Tohidi, M.; Ghasemi-Varnamkhasti, M.; Ghafarinia, V.; Mohtasebi, S.S.; Bonyadian, M. Identification of trace amounts of detergent powder in raw milk using a customized low-cost artificial olfactory system: A novel method. *Measurement* **2018**, *124*, 120–129. [CrossRef]