Università degli Studi di Genova

Doctorate in Sciences and Technologies of Chemistry and Materials

(XXXV cycle)

Curriculum: Pharmaceutical, Food and Cosmetic Sciences

DIFAR - Department of Pharmacy, University of Genova, Viale Cembrano 4, 16148 - Genova

Development of innovative analytical methods based on spectroscopic techniques and multivariate statistical analysis for quality control in the food and pharmaceutical fields.

ELEONORA MUSTORGI

CONTENTS

ABSTRACT

The increasing demand on quality assurance and ever more stringent regulations in food and pharmaceutical fields are promoting the need for analytical techniques enabling to provide reliable and accurate results. However, traditional analytical methods are labor-intensive, time-consuming, expensive and they usually require skilled personnel for performing the analysis. For these reasons, in the last decades, quality control protocols based on the employment of spectroscopic methods have been developed for many different application fields, including pharmaceutical and food ones. Vibrational spectroscopic techniques can be an adequate alternative for acquiring both chemical and physical information related to homogenous and heterogenous matrices of interest. Moreover, the significant development of powerful data-driven methodologies allowed to develop algorithms for the optimal extraction and processing of the complex spectroscopic signals allowing to apply combined approaches for quantitative and qualitative purposes.

The present Doctoral Thesis has been focused on the development of ad-hoc analytical strategies based on the application of spectroscopic techniques coupled with multivariate data analysis approaches for providing alternative analytical protocols for quality control in food and pharmaceutical sectors.

Regarding applications in food sector, excitation-emission Fluorescence Spectroscopy, Near Infrared Spectroscopy (NIRS) and NIR Hyperspectral Imaging (HSI) have been tested for solving analytical issues of independent case-studies. Unsupervised approaches based on Principal Component Analysis (PCA) and Parallel Factor Analysis (PARAFAC) have been applied on fluorescence data for characterizing green tea samples, while quantitative predictive approaches as Partial Least Squares regression have been used to correlate NIR spectra with quality parameters of extra-virgin olive oil samples. HSI was applied to study dynamic chemical processes which occur during cheese ripening with the aim to map chemical and sensory changes over time.

The rapid technical progress in terms of spectroscopic instrumentations has led to have more flexible portable systems suitable for performing measurements directly in the field or in a manufacturing plant. Within this scenario, NIR spectroscopy proved to be one of the most powerful Process Analytical Technologies (PAT) for monitoring and controlling complex manufacturing processes. In this thesis, two applications based on the implementation of miniaturized NIR sensors have been performed for the real-time powder blending monitoring of pharmaceutical and food formulation, respectively. The main challenges in blending monitoring are related to the assessment of the homogeneity of multicomponent formulations, which is crucial to ensure the safety and effectiveness of a solid pharmaceutical formulation or the quality of a food product. In the third chapter of this thesis, tailor made qualitative chemometric strategies for obtaining a global understanding of blending processes and to optimize the endpoint detection are presented.

RIASSUNTO

Il maggior interesse da parte dei consumatori rispetto al controllo qualità di un prodotto finito e l'entrata in vigore di normative sempre più stringenti, nei settori alimentare e farmaceutico, hanno spinto la ricerca verso lo sviluppo di tecniche analitiche che consentano di fornire risultati affidabili e accurati. Tuttavia, i metodi analitici tradizionali sono laboriosi, dispendiosi in termini di tempo, costosi e di solito richiedono personale qualificato per eseguire le analisi. Per questi motivi, negli ultimi decenni, sono stati sviluppati protocolli di controllo qualità basati sull'impiego di metodi spettroscopici in diversi campi di applicazione, inclusi quelli farmaceutico ed alimentare. Le tecniche spettroscopiche vibrazionali possono essere un'alternativa adeguata per l'acquisizione di informazioni sia chimiche che fisiche relative a matrici di interesse siano esse omogenee o eterogenee. Inoltre, il significativo sviluppo di strategie multivariate di analisi dei dati ha permesso di sviluppare algoritmi per l'estrazione e l'elaborazione dell'informazione spettrale consentendo di calcolare modelli predittivi qualitativi e quantitativi.

La presente tesi di dottorato è stata incentrata sullo sviluppo di strategie analitiche ad-hoc basate sull'applicazione di tecniche spettroscopiche accoppiate con approcci di analisi di dati multivariati per fornire protocolli analitici alternativi per il controllo di qualità nei settori alimentare e farmaceutico.

Per quanto riguarda le applicazioni nel settore alimentare, la spettroscopia di fluorescenza ad emissione di eccitazione, la spettroscopia nel vicino infrarosso (NIRS) e l'imaging iperspettrale NIR (HSI) sono state testate su casi studio indipendenti. Approcci esplorativi basati sull'analisi delle componenti principali (PCA) e sull'Analisi Fattoriale Parallela (PARAFAC) sono stati applicati su dati di fluorescenza per caratterizzare campioni di tè verde, mentre approcci predittivi quantitativi come la regressione dei minimi quadrati parziali sono stati utilizzati per correlare gli spettri NIR con i parametri di qualità di campioni di olio extra- vergine di oliva. L'Imaging Iperspettrale è stato invece applicato per studiare i processi biochimici che si verificano durante la maturazione del formaggio con l'obiettivo di mappare i cambiamenti chimici e sensoriali nel tempo.

Il rapido progresso tecnico in termini di strumentazioni spettroscopiche ha portato ad avere sistemi portatili più flessibili adatti ad effettuare misure direttamente sul campo o in un impianto produttivo. All'interno di questo scenario, la spettroscopia NIR si è rivelata una delle tecnologie analitiche di processo (PAT) più potenti per il monitoraggio e il controllo di processi di produzione complessi. In questa tesi, si riportano due diverse applicazioni basate sull'implementazione di sensori NIR miniaturizzati per il monitoraggio in tempo reale della miscelazione di polveri rispettivamente di formulazioni farmaceutiche e alimentari. Le principali sfide nel monitoraggio della miscelazione sono legate alla valutazione dell'omogeneità di formulazioni complesse; l'uniformità della formulazione è infatti fondamentale per garantire la sicurezza e l'efficacia di una formulazione farmaceutica solida o la qualità di un prodotto alimentare. Nel terzo capitolo di questa tesi, vengono presentate strategie chemiometriche qualitative su misura per incrementare la conoscenza di un processo di miscelazione su scala industriale al fine di ottimizzare il rilevamento del punto finale della miscelazione.

PREFACE

In the last decades, the interest in rapid, non-destructive, and accurate analytical methods for quality control in food and pharmaceutical fields has been continuously increasing. In the literature, many studies demonstrated the suitability of vibrational spectroscopic techniques in measuring physicochemical properties of samples in a fast way and without requiring time-consuming sample preparation. To solve the emerging challenges in agrifood and pharmaceutical sectors, during my PhD, I tested two vibrational spectroscopic methods: Excitation-Emission fluorescence spectroscopy and Near Infrared (NIR) spectroscopy; this latter has been investigated using FT-NIR benchtop instruments, miniaturized NIR sensors and hyperspectral imaging systems. The significant volume of data generated by these techniques require the application of chemometric strategies to extract the useful information for the calculation of qualitative and quantitative models. The main goal of my thesis was to develop ad-hoc chemometric strategies for modeling complex spectroscopic signals with the aim to define innovative analytical protocols for challenging applications in food and pharmaceutical fields.

The thesis is organized in four chapters, one for each spectroscopic technique tested, and a final chapter which includes the overall conclusion for a comprehensive evaluation of the scientific impacts of the research work performed during my PhD.

Chapter 1 provides an overview of Excitation-Emission Fluorescence spectroscopy with a focus on the chemometric approaches applied for processing three-way Excitation- Emission matrices (EEMs). The two paragraphs of this chapter report two independent case-studies in which Parallel

Factor Analysis (PARAFAC) has been applied on fluorescence data for developing solutions for agrifood and biomedical fields, respectively.

Chapter 2 is related to Near Infrared Spectroscopy for the analysis of food products with a focus on the application of traditional benchtop instrumentation. In more detail, Partial Least Square (PLS) regression models have been developed for correlating the NIR spectra of Extra Virgin Olive Oils (EVOO) with key quality parameters in order to statistically compare the analytical performance of quartz cuvettes with disposable glass vials. In the second paragraph, an innovative approach, based on the application of Aquaphotomics, has been developed to investigate changes in water molecular structure during the storage of rice germ samples.

Chapter 3 illustrates the potential of the online implementation of Process Analytical Technology (PAT) systems, based on NIR sensors, for the online monitoring of complex manufacturing processes in food and pharmaceutical fields, such as powder blending. Multivariate Statistical Process Control (MSPC) models have been developed for outlining a strategy to process NIR signals acquired along the mixing of a zootechnical formulation, with the aim to optimize the endpoint detection. During my internship at the University of Barcelona, I had the possibility to test on a real industrial case-study different unsupervised qualitative approaches for the monitoring of a blending process. In more detail, I focused on the development of chemometric approaches based on the application of Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) for studying the evolution of low dosage formulations during mixing. The second paragraph shows a critical comparison of different multivariate strategies, including the Moving Block F-test, MSPC based on PCA and MCR-ALS.

Chapter 4 is dedicated to Hyperspectral Imaging (HSI). The higher quality standards and the growing awareness of customers in the food sector led to find more advanced analytical techniques for the chemical-physical characterization of food products. HSI allows the analyst to obtain both spectral and spatial

information from a non-destructive analysis of the sample. In this way, it is possible to map and follow dynamic processes which occur in complex food matrices over time. In this chapter, an innovative analytical approach, based on hyperspectral imaging in the near-infrared region (HSI-NIR) and multivariate pattern recognition, to study and monitor the extent – spatial and temporal – of biochemical phenomena responsible for cheese ripening is reported.

Chapter 5 shows the main conclusions resulting from the present work. At the end of this chapter the list of publications and details about other activities such as the conference participations and the courses attended during the three years have been reported.

CHAPTER 1: FLUORESCENCE SPECTROSCOPY

Fluorescence spectroscopy deals with excitation and emission in molecules. The absorption of light by electrons, that occupy specific orbital in a population of molecules, can elevate one electron to an upper vacant orbital having higher energy; in this way excited states are produced. While the absorbance is only related to the transition from ground state to excited state, the fluorescence involves the relaxation from excited to ground state. Only a certain number of molecules (generically called fluorophores) that usually present aromatic rings, conjugated double bonds or other similar rigid structures, can emit energy in the form of fluorescence returning to the ground state [1].

When the radiation is absorbed by the molecule, an electron is elevated from the ground singlet states, S_0, to an excited singlet state, S_1 and the molecule is transferred to an electronically excited state. After excitation, the molecule will undergo a rapid internal conversion to the lowest vibrational level of the excited electronic state prior to emission. Finally, fluorescence emission occurs when the molecule returns to the more stable ground state S_0; the fluorescent radiation is emitted at a wavelength which depends on the difference in energy between the two electronic states [2].

Using ultraviolet or visible light it is possible to promote the fluorophore of interest to one of several vibrational levels for the given electronically excited level. This means that absorption and fluorescence emission can occur over a broad range of wavelengths, describing the detailed fluorescence characteristics of molecules. Therefore, it follows that an emission spectrum is measured as the radiation emitted across a broad wavelength range upon excitation at a selected wavelength. Similarly, the excitation spectrum can be measured by fixing the emission at one fixed wavelength while exciting the molecule over a wavelength range. When measuring several emission spectra over a range of shifting excitation wavelengths, it's possible to obtain an excitation emission matrix (EEM). EEMs consist of three order arrays (sample × excitation wavelength × emission wavelength) that require proper multi way method to extract the useful information. This enables determination of the number of

fluorophores in the multicomponent system and the extraction of their excitation and emission spectra. One of the most applied chemometric multi-way methods for handling fluorescence data is PARAFAC (PARAllel FACtor analysis) [3]. This method allows to decompose the EEM into trilinear terms and a residual array in according to the number of fluorophores detected in the samples. When data are accurately modeled, the obtained parameters of the model can be further used for calculation of the relative concentration of fluorophores in samples. PARAFAC minimizes the sum of squares of the residual e_{ijk} using a least squares algorithm, as shown in the following equation:

$$x_{ijk} = \sum_{f=1}^{F} a_{if}\, b_{jf}\, c_{kf} + e_{ijk} \qquad \text{i=1,2,…,I; j=1,2,…,J; k=1,2,…,K}$$

where x_{ijk} is the original fluorescence data matrix for sample *i,* excitation wavelength *j* (mode 2) and emission wavelength *k* (mode 3) and e_{ijk} is the residual matrix which represents the variance not explained by the model. In this way, from the original three-way fluorescence data array, it's possible to obtain a set of sample scores a_{if}, loadings for the excitation mode b_{if}, and loadings for the emission mode c_{kf} for each component *f*. A robust procedure to validate a PARAFAC model is through the application of the split half analysis [4], which divides the initial data set into two halves and calculating two independent PARAFAC models. Since the solution of PARAFAC has to be unique, if the correct number of components has been selected, both models will provide the same result. Moreover, in order to determine the proper number of components of a PARAFAC model, it's also possible to calculate the core consistency [5] and the percentage of explained variance. If the model is completely trilinear the core consistency will be equal to 100%.

In recent years, combined approaches of excitation-emission fluorescence spectroscopy coupled with PARAFAC have gained wide acceptance in different sectors as in chemical, agri-food, pharmaceutical, environmental, and clinical one.

In this project, the potential advantages in terms of sensitivity and specificity of the present method have been tested for developing ad-hoc solutions for two different case studies.

In the first paragraph of this chapter, an alternative analytical approach for the chemical characterization of green tea (GT) samples in according to geographical origins has been proposed. The application of excitation–emission fluorescence spectroscopy coupled with multivariate data analysis algorithms as Principal Component Analysis (PCA) [6] and Partial Least Squares Class-Modelling (PLS-CM) [7] allowed to recognize and classify properly Japanese and Chinese green tea samples. However, in order to provide a meaningful chemical explanation about these differences, PARAFAC has been applied showing a clearer correlation between the geographical origins of the samples and the content of antioxidant compounds, especially catechins.

In the second paragraph, a challenging biomedical application of this analytical approach has been reported. Prostate cancer is the second most widespread malignant tumor in the male population and based on the latest scientific evidence, it is of current interest to have rapid and accurate analytical methods for early screening of prostate cancer directly by urine analysis, in order to provide reliable results while improving patient compliance. Thanks to the collaboration with the University of Pisa, (Urology Department), it was possible to analyze a total of 69 urine samples (46 samples from patients with histologically proven prostate cancer and 23 from healthy donors) using a Perkin-Elmer LS55B luminescence spectrometer. The application of PARAFAC allowed resolution of the spectral profiles corresponding to single fluorophores and their relative concentration estimation, which were then fed to discriminant classifiers that allowed to develop a first attempt of healthy/cancer discrimination model. This analytical approach can contribute in defining a simple and non-invasive protocol for prostate cancer detection as a screening tool able to support traditional diagnostic methods.

1.1 AN ALTERNATIVE ANALYTICAL APPROACH FOR THE CHARACTERIZATION OF GREEN TEA BASED ON EXCITATION-EMISSION FLUORESCENCE SPECTROSCOPY AND PARALLEL FACTOR ANALYSIS (PARAFAC)

Scientific Background and aim of the work

Tea is an aromatic beverage made from the leaves of *Camellia sinensis*, a plant native to Southeast Asia, cultivated and consumed by humans for thousands of years. Due to its attractive aroma and taste and its effect on reducing lifestyle-related diseases, tea is the most consumed beverage in the world. Green tea (GT) is made from unfermented leaves of *Camellia sinensis* and contains a high concentration of polyphenols, which are powerful antioxidants. The potential health benefits of GT, especially related to its antioxidant properties, have led to an increase of its consumption in the last decades. The principal compounds of GT having biological effects have been identified as catechins and xanthines [8]. Catechins show a strong antioxidant activity and exert antiinflammatory, antiarhtritic, antiangiogenic, neuroprotective, anticancer, antiobesity, antiatherosclerotic, anti-diabetic, antibacterial, antiviral and antidental caries effects. Xanthines are responsible for the stimulating effects; caffeine (CF) is a central nervous system and cardiac stimulant and has a diuretic effect, while theobromine (TB), which is present in lower amounts, has also a diuretic effect [8], [9], [10], [11], [12], [13]. Among the most abundant catechins in GT there are (+)-catechin, ((+)C), (-)-epicatechin (EC), (-)-epigallocatechin (EGC), (-)-epicatechingallate (ECG), (-)-epigallocatechin gallate (EGCG) [14].

The composition of GT can be influenced by several parameters associated with growth conditions, such as genetic strain, season, climatic conditions, soil profile, growth altitude, horticultural practices, plucking season, shade growth, and with the region in which tea has been cultivated. The other factors that can influence the profile

of bioactive compounds are manufacturing process (withering, steaming/pan-firing, rolling, oxidation/fermentation and drying) and storage [15,16]. Besides this huge variability, the price of tea greatly varies according to its geographical origin. Hence, the recognition of the origin of GT is crucial to protect the interests of both consumers and sellers [17,18]. Several analytical methods have been proposed together with chemometric techniques in order to characterize the geographical origins and/or varieties of teas [19], [20], [21], [22].However, most of these methods require expensive equipment and involve tedious sample preparation in order to discriminate GT samples from different geographical origins; as an example, Ye et al. [21] extracted the volatile organic components from the dried tea leaves by headspace solid-phase microextraction procedure, followed by GC–MS analysis.

In a previous paper coauthored by a colleague from the research group with whom we collaborate [17], cyclodextrin-modified micellar electrokinetic chromatography (CyD-MEKC) was employed to simultaneously analyse the most represented catechins and methylxanthines in 92 GT samples of different geographical origin, and the comparison of the obtained data showed that Japanese commercial GT products contained a general lower level of catechins than Chinese GTs. The contents of catechins and methylxanthines were thus used as chemical descriptors and potential indicators of the geographical origin. Considering this previous work as a starting point for further investigations, in the present study an alternative analytical approach was applied for identifying the differences in terms of active compounds content in GT samples from different geographical origin. In order to reach this aim, 63 GT samples were analysed by fluorescence spectroscopy: 29 samples from Japan and 34 from China. The main reason of the choice of these two countries was the interest of the consumers in the comparison of Japanese and Chinese GTs in terms of active compounds content. As a matter of facts, Chinese GT tends to cost consumers much less than Japanese GT, for the massive prevalence of Chinese GT and thus the necessity of maintaining low prices by Chinese producers, and for the lack of space for the production of GT in Japan. Moreover, one of the main differences in GT processing between Chinese and Japanese producers is the way deactivation of enzymes is performed. Chinese GT is usually dry heated in order to deactivate oxidases, whereas in the case of Japanese GT steaming is employed. Besides,

Japanese GT is usually shade grown [16]. Hence, we deemed it worthwhile to compare the GTs from these two countries in order to understand if the higher price of Japanese teas can be supported or not by the fact that it is a more prized tea for its higher antioxidant capacity.

In more detail, the innovative analytical approach presented is based on the combination of excitation–emission fluorescence spectroscopy and chemometric tools to extract useful information from a huge amount of data. The chemometric approach is a fundamental part of the interpretation of fluorescence spectral data of agro-food products due to the presence of many fluorophores, since the fluorescence of a sample consists of a number of overlapping signals not easily understandable without a proper data processing. Accordingly, to these principles, three-dimensional fluorescence spectra were elaborated through PCA [5] after unfolding the data into matrices and through Parallel Factor Analysis (PARAFAC) [4] on three-way data as display methods. Moreover, SELECT [18] technique was applied for variable selection, in order to individuate the variables with the highest classification power, i.e. the most informative emission bands in discriminating between Japanese and Chinese GTs.

Finally, the content of catechins and methylxanthines was determined in a subset of 24 GT samples by the previously developed chiral CyD-MEKC method in order to obtain complementary information on the geographical origin of GT samples and to confirm what observed in the previous work [17], i.e. that the amount of all the considered compounds was higher for Chinese GTs, with the exception of ECG. A Partial Least Squares Class-Modelling (PLS-CM) [7] was carried out on this subset of samples to develop a predictive model able to classify new GT samples according to the geographical origin using the CyD-MEKC data.

Experimental plan: sampling, chemical and spectroscopic analysis

The reference standards of (+)C, EC, EGC, ECG, EGCG, CF, TB, as well as boric acid, 86.1% phosphoric acid, sodium dodecyl sulphate (SDS), (2-hydroxypropyl)-β-cyclodextrin (HPβCyD, degree of substitution 0.6), were purchased from Sigma-Aldrich (St. Louis, MO, USA). The standard stock solutions (1 mg mL^{-1}) of (+)C, EC, EGC, ECG, EGCG, CF, TB and of the internal standard syringic acid were prepared in a mixture of methanol/water in 15:85 ratio %v/v. Working standard solutions were obtained by dilution with water in a vial to 500 µL for achieving the desired final concentration values of the compounds.

A set of 63 GT samples of different varieties and from different geographical origins (29 from Japan and 34 from China) was selected for the study and analysis. In order to assure a good degree of representativity of the samples, the main sources of variability for GTs were considered, *i.e.* for Japanese GTs the different varieties, including Bancha, Gyokuro, Matcha, Sencha, Matcha Tsuru types, while for Chinese GTs the different zones (the ten provinces of Hunan, Fujian, Zhejiang, Anhui, Yunnan, Guandong, Jiangsu, Hubei, Shandong, Guanxi). Moreover, each geographical group included samples stored in different conditions and coming from different manufacturing processes. The commercial GT samples were collected locally in specialized stores located in the cities of Florence and Genoa (Italy). A subset of 24 samples randomly selected including different types of Japanese GT and different zones of Chinese GT has been analyzed using the CyD-MEKC method for the quantitation of catechins and methylxanthines (Table 1).

Table 1: GT samples analysed by the CyD-MEKC method: content of catechins and methylxanthines[a]

Sample ID[b]	Category[c]	EC	ECG	EGC	CF	ECGC	(+)C	TB
J1	1	8.64	16.07	6.03	13.08	12.08	0.14	0.05
J2	1	6.82	16.24	4.35	15.13	11.6	0.25	0.09
J3	1	7.02	13.81	7.96	16.79	14.71	0.27	0.23
J6	1	8.94	14.44	7.71	9.95	11.46	0.33	0.93
J8	1	6.93	15.33	8.23	14.64	16.21	0.15	0.21
J9	1	0.76	1.22	0.89	6.1	2.2	0.22	0.24

J12	1	0.79	1.23	0.99	5.9	2.08	0.16	0.29
J13	1	0.38	1.21	1.35	8.13	3.09	0.23	0.22
J17	1	1.92	5.01	2.09	5.39	4.08	0.08	0.04
J23	1	7.1	14.13	5.64	16.95	12.11	0.17	0.15
J24	1	6.97	46.4	4.32	14.98	11.51	0.25	0.12
J29	1	7.05	14.67	5.28	14.36	12.02	0.22	0.13
C1	2	6.09	10.46	14.66	11.72	14.32	1.39	3.17
C2	2	5.77	4.29	23.12	23.38	18.38	1.53	1.46
C4	2	4.71	6.65	21.37	15.49	12.42	0.24	1.68
C6	2	15.86	10.61	38.93	35.95	27	3.24	2.42
C7	2	7.66	6.29	8.44	21.82	12.68	0	0.92
C8	2	6.47	14.88	32.69	20.84	19.93	0.63	2.28
C10	2	7.03	6.65	23.57	32.26	30.89	1.55	3.07
C12	2	5.8	8.05	6.32	19.69	12.15	0	0.64
C13	2	5.03	7.12	7.49	19.37	13.3	0.39	1.16
C14	2	4.52	5.39	7.64	18.54	14.77	0.44	1.59
C16	2	10.19	8	23.28	27.24	20.88	1.84	2.01
C22	2	4.87	3.45	14.44	16.27	11.34	0.3	0.32

[a] The data are expressed as the average content in mg g^{-1}, dry basis (mean of two determinations).
[b] Sample code.
[c] Category 1: Japanese GT samples; category 2: Chinese GT samples.

The CyD-MEKC method used for the determination of the compounds was derived from the previous study [22]. The analyses were carried out using a 3DCE instrument from Agilent Technologies (Waldbronn, Germany) controlled by the software 3DCE ChemStation (Agilent Technologies) for both acquisition and data management. Fused-silica capillaries (Unifibre, Settimo Milanese, Italy) of 33.0 total length, 8.5 cm effective length and 50 μm inner diameter were used. The detection was carried out by using the on-line DAD detector and the detection wavelength was 200 nm. Voltage and temperature were set at 15 kV and 25 °C, respectively. The background electrolyte was made by 25 mM borate-phosphate buffer pH 2.50 with the addition of 90 mM sodium dodecyl sulphate and 25 mM HPβCyD. Total analysis time was about 8 min. Calibration was performed by the internal standard method, using syringic acid as internal standard. The method had been previously validated in terms of selectivity, linearity, repeatability, accuracy and sensitivity, showing adequate performances for the analysis of catechins and methylxanthines in GT, with LOQ values ranging from 0.05 to 0.7 μg mL−1 [22]. Further information on the CE method and procedure may

be found in mentioned Ref.[22]. The EEM fluorescence measurements were performed directly on GT extracts at room temperature on a Perkin-Elmer LS55B luminescence spectrometer (Waltham, MA, USA). The excitation-emission matrices of the GT infusions were recorded using the standard cell holder and a 10 mm quartz SUPRASIL® cell with cell volume of 3.5 mL by PerkinElmer. The excitation spectra were recorded between 200 nm and 290 nm each 5 nm (19 recorded points), whereas the emission wavelengths ranged from 295 nm to 800 nm each 0.5 nm (1011 recorded points). The excitation and the emission monochromator slits were set to 10 nm. The FL WinLab software (PerkinElmer) was used to register the fluorescent signals.

Chemometric approach purposed

In this project, the data processing strategy applied for handling the EEM data, included three different steps:

a) *Data Exploration*: Unsupervised exploratory techniques such as PCA and PARAFAC have been applied to obtain a global understanding of the system.

b) *Variable selection*: SELECT has been applied for identifying the most significant spectroscopic variables for characterizing the samples.

c) *Class modeling*: PLS-CM has been used for developing a predictive model for classifying properly GT samples in according to the geographical origin.

a) Data Exploration

PCA [4] is the most used tool in exploratory data analysis and it uses an orthogonal transformation to convert a set of correlated variables into a set of uncorrelated variables called principal components. This approach makes it possible to visualize in a comprehensive way the dataset starting from a two-dimensional data matrix. According to the specific nature of EEM data, organized in a three-dimensional data array, for performing PCA a step of unfolding of the matrix is requested, while with the PARAFAC algorithm it is possible to directly model n-way data. In the case of three-way data, like the EEM data, PARAFAC decomposes a data array **X** with dimension $I \times J \times K$ into three loading matrices **A**, **B** and **C,** being their columns a_i, b_j and

c_k respectively (see the PARAFAC description in the Introduction of Chapter 1). The trilinear PARAFAC model is expressed as follows:(1)$x_{ijk}=\sum_{f=1}^{F} a_{if} b_{jf} c_{kf}$ i=1,2,…,I; j=1,2,…,J;k=1,2,…,K where x_{ijk} is the element in the position *i*, *j*, *k* of the three-way array **X**; *F* is the number of factors; a_{if}, b_{jf} and c_{kf} are the elements of the matrices **A** (I × F), **B** (J × F) and **C** (K × F), respectively; e_{ijk} represents the generic element of the residual array **E** (I × J × K). The PARAFAC model is found by minimizing the sum of squares of the residuals.

The excitation-emission fluorescence matrices obtained for several samples can be arranged into a three-way array and the PARAFAC decomposition can be applied for the analysis of fluorescent data. In this case, **X** contains the fluorescence intensity at the *k*-th excitation wavelength and *j*-th emission wavelength recorded for the *i*-th sample. Therefore, the vectors a_i, b_j and c_k are the sample, emission and excitation profiles of the *f*-th fluorophore, respectively. The similarity between the trilinear PARAFAC model and the physical model for fluorescence can be found in Ref. [23].

The core consistency diagnostic (CORCONDIA) developed by Bro and Kiers [5] is an index that measures the degree of trilinearity of the experimental data array. A trilinear model has a value of CORCONDIA index close to 100%.

If the fluorescence data are trilinear and the appropriate number of factors has been chosen to fit the model, the PARAFAC decomposition provides unique profile estimations, and the achievement of the true underlying excitation and emission spectra for every fluorophore is ensured [3]. PARAFAC has been widely used due to this highly attractive uniqueness property [23], which could be used for the unequivocal identification of compounds.

b) Variable selection

The selection of the informative variables was performed by means of SELECT [24], a feature selection technique based on the stepwise decorrelation of the variables, which is implemented in the chemometric V-Parvus software [25]. This technique generates a set of decorrelated variables ordered according to their Fisher weights. At each step, SELECT searches for the variable with the largest classification weight.

This variable is selected and decorrelated from the other variables; then the algorithm is repeated until a fixed number of variables is selected, or the Fisher weight is lower than a specific cut-off value. SELECT presents an interesting characteristic: the fraction of the residual variance of the predictors after the orthogonalization can be used to select intervals of predictors with better classification performance.

c) Class modeling

PLS-CM [7] is a supervised method of classification between two categories (or classes), in our case Japanese or Chinese GT. It is a version of Partial Least Squares (PLS) algorithm with a binary response that makes it possible to model the probability distribution of the samples for each class and then performs a hypothesis test evaluating the α probability of type I error and the β probability of type II error. Class-model sensitivity (proportion of the samples of the class that are correctly assigned) and specificity (proportion of samples correctly rejected) are (1-α)·100 and (1-β)·100, respectively. The risk curve is the plot of β error *versus* α error probabilities. Data analysis was performed in the MATLAB environment [26], thanks to tailor made algorithms developed and implemented by the Authors. For the data processing, PCA, PARAFAC and PLS-CM algorithms were applied, in order to extract the significant information embodied within data. For performing variable selection, the SELECT method was applied thanks to its implementation in the software V-Parvus [25].

Research outcomes

The CyD-MEKC method previously described [22] was applied to the analysis of a subset of 24 GT samples in order to confirm our previous observations [17] and to lay the basis for the EEM data processing. By applying the CyD-MEKC method, the samples were characterized by means of $n = 7$ variables, namely (+)C, EC, EGC, ECG, EGCG, CF and TB (mg g^{-1}, dry basis), obtaining a data matrix having 24 rows (samples) and 7 columns (variables), shown in Table 1. This data set was submitted to chemometric modeling starting from PCA as a display method and then applying the PLS-CM algorithm for class modeling purposes.

Firstly, PCA was performed on the data matrix to enhance the presence of structures inside the samples and to understand the correlation between the variables. Fig. 1 shows the loading (a) and the score (b) plots of the catechins ((+)C, EC, EGC, ECG, EGCG), CF and TB autoscaled data in the plane of the 2 first Principal Components, that explain the 86% of the total variance. From the loading plot it was possible to point out that the variable EGCG is the most important factor in PC1, followed by CF and EGC. All loadings are positive so that the samples with highest scores on PC1 have greater value in all the variables. On the contrary, loadings of PC2 have different sign: ECG has the highest positive loading and TB has the highest negative. Along PC1, the scores of the Japanese GT samples in relation to the scores of the Chinese GT samples are lower, indicating that in general Chinese GT samples were characterized by a higher content in the active compounds. This observation is in full agreement with what reported in our previous study [17].

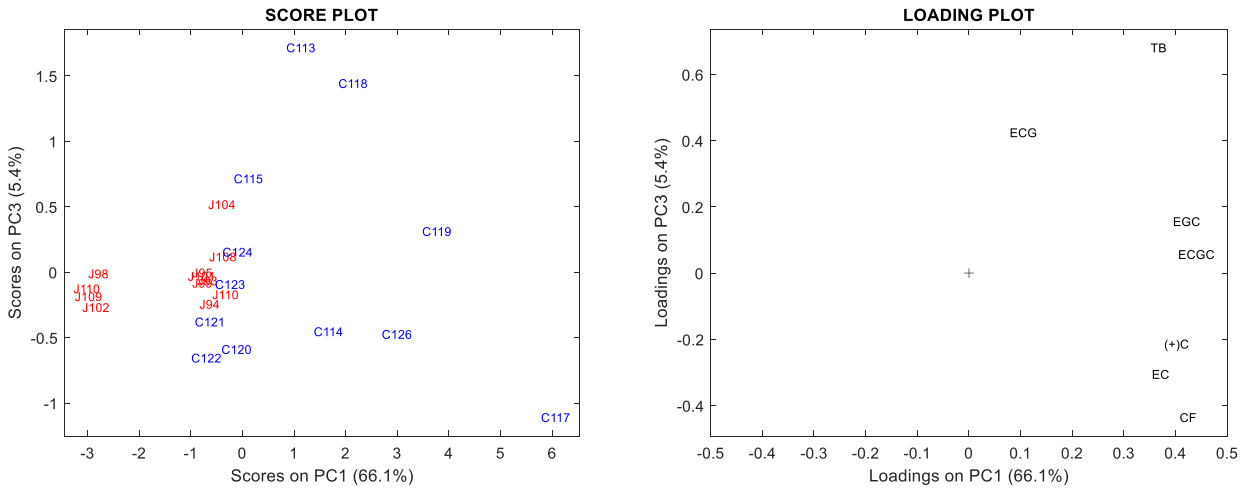

Figure 1:PCA (on the left) score plot and (on the right) loading plot of catechins and methylxanthines data

In order to build the PLS-CM model, it is necessary to build a dummy vector containing the information about class membership; for this reason, a binary response was

constructed considering the values 1 and 2 for the Japanese and Chinese GT, respectively (Table 1). The number of PLS latent variables that minimized the root mean square error in cross-validation (RMSECV) obtained by leave one out procedure was 3, and they explained the 81.68% of response with 90.05% of predictors variance. Fig. 2 shows the distribution of PLS fitted values for the Japanese and Chinese GT samples. Both classes have normal distribution with mean values 1.09 and 1.91 and Standard Deviation values 0.09 and 0.27, respectively.

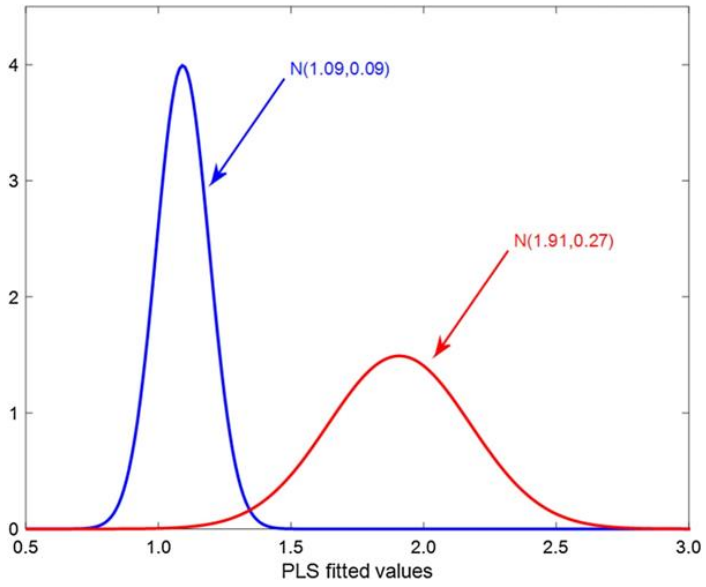

Figure 2: Normal distribution fitted for Japanese GT samples (in blue, on the left) and Chinese GT samples (in red, on the right).

In order to decide if an unknown sample belongs to one or another class, a threshold value, tv, between 1 (GT from Japan) and 2 (GT from China) must be established. If the value estimated by PLS is higher than tv the sample is classified to belong to class 2 (China), while for estimated values lower than tv the sample is classified to belong to class 1 (Japan). A model for one class (e.g. “GT Japanese”), is in fact the acceptation region for the null hypothesis H_0: the sample belongs to “Japanese GT” class. Therefore, the evaluation of the quality of a class model is given by its sensitivity and specificity. Both parameters have been evaluated in cross-validation, being 98.70% and 98.68%, respectively.

Fig. 3 shows two typical excitation-emission spectra of one Japanese (J1) and one Chinese GT sample (C1).

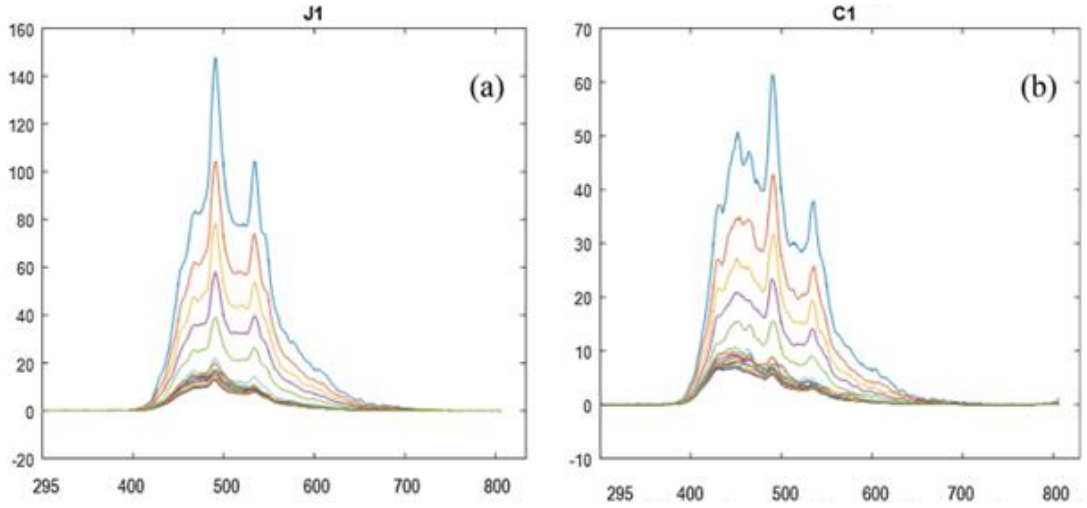

Figure 3: A typical excitation-emission spectra of (a) a Japanese (J1) and (b) a Chinese (C1) GT sample.

In order to assess the experimental variability and the repeatability in preparing the tea infusions, the analysis of two GT samples of different geographical origin (one from Japan and one from China) were replicated 3 times at a distance of time (one week). PC1, which explains 97.8% of the total variance, clearly separates the 2 GT samples; on the contrary, the difference among the 3 replicates of the same sample is along PC2, which explains only 1.4% of the variance (data not shown).

a) <u>Data Exploration: PCA results</u>

Two bands of the emission spectra were removed, namely from 295 to 350 nm and from 700 to 800 nm, due to the lack of information typical of these two areas (Fig. 3). The range between 350–700 nm was retained and used for data elaboration. A data matrix of dimension 63 × 13,300 was built, where each row corresponded to the emission spectrum (700 wavelengths) obtained at each of the 19 excitation wavelengths for all the 63 GT samples measured. PCA was performed as

unsupervised pattern recognition technique on this 'unfolded' matrix after the data had been mean-centered. Fig. 4 shows the score plot on the plane PC1-PC4. It is possible to notice a discrimination between Japanese and Chinese GT samples along PC1, the direction explaining the 74.3% of the total variance, even if a certain overlap is present and the complete separation between the classes is not obtained. In the PC1-PC4 plot it can be also clearly noticed that Matcha GT samples, considered one of the Japan's rarest and most precious GT variety, are grouped in a cluster in the orthogonal space at negative scores on PC1.

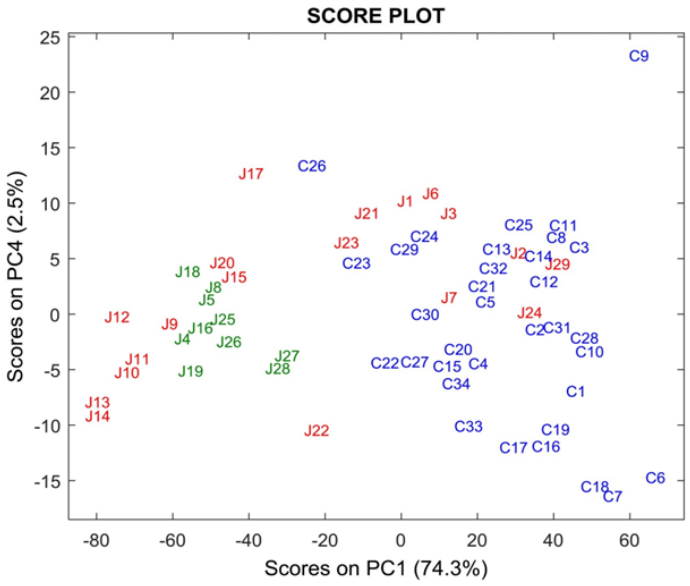

Figure 4: PCA score plot on the PC1-PC4 plane for the fluorescence data. Matcha samples are indicated in green in the plot

Looking at the loading profile on PC1 (Fig. 5), it is possible to notice the bands more informative along PC1 and thus useful for separating between Japanese and Chinese GTs, namely 410–450 nm and 500–600 nm. The first band (410–450 nm) shows positive loadings on PC1 and this suggests that it is related to active compounds content in GT from China; on the contrary the broad band (500–600 nm) has negative loadings, therefore it seems linked to chemical compounds characterizing the Japanese GTs.

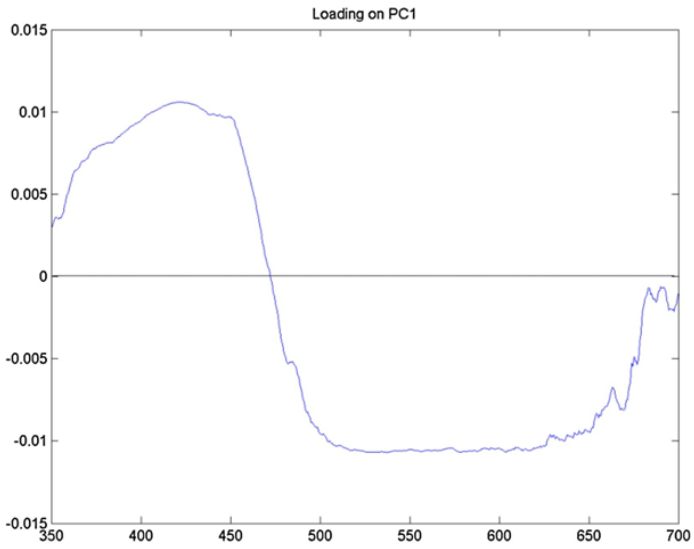

Figure 5: Loading profile on PC1.

a) *Data exploration: PARAFAC results*

The EEM data recorded for the 63 samples analyzed were arranged into a three-way data array where the excitation wavelengths between 200 nm and 290 nm and the emission wavelengths between 295 nm and 800 nm were considered. Therefore, the dimension of this array was 63 × 1011 × 19 (where 63 are the samples, 1011 the emission wavelengths and 19 the excitation wavelengths). The PARAFAC decomposition of this array, without any constraint, required two factors (CORCONDIA of 100%, explained variance of 98.6%). The plot of the loadings of the mode of the samples (first mode, Fig. 6a) is similar to the PCA score plot (Fig. 4) and it shows a rather clear discrimination between Chinese and Japanese GTs. The plot of the loadings of the mode of the emission (second mode, Fig. 6b) shows the emission spectra for two fluorophores, one with maximum around 420 nm and the other one with maxima at 500–550 nm. As can be seen in these plots, PARAFAC enabled to differentiate the infusions of GT according to the geographical origin (Chinese and Japanese). Moreover, due to the trilinearity of the data, it can be concluded that the two groups of fluorophores found with the PARAFAC model are the same in all the GT samples.

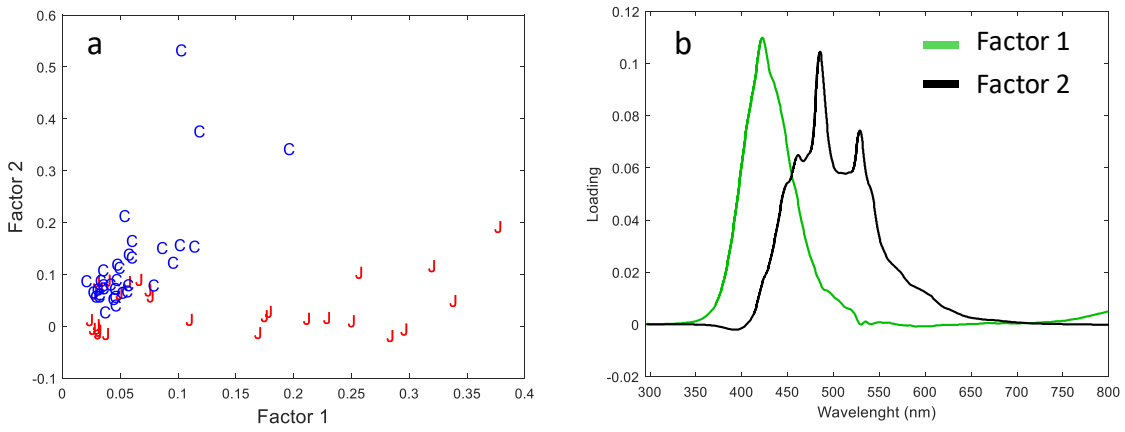

Figure 6:PARAFAC results: a. loading plot of the mode of the samples (first mode); explained variance 98.6% (F1 = 96.0% and F2 = 2.6%); b.the loading plot of the emission mode (second mode)

b) Variable selection: SELECT outcomes

SELECT was applied as a variable selection technique in order to individuate the variables with the highest classification power, i.e. the most informative emission bands in discriminating between Japanese and Chinese GT samples. SELECT was applied on the unfolded data matrix of dimension 63 × 13,300 where each row corresponded to the emission spectrum obtained for each excitation wavelength of each GT sample measured; the frequency histogram of the selections (Figure 7) showed as the most selected variables the two bands 415–450 nm and 495–550 nm.

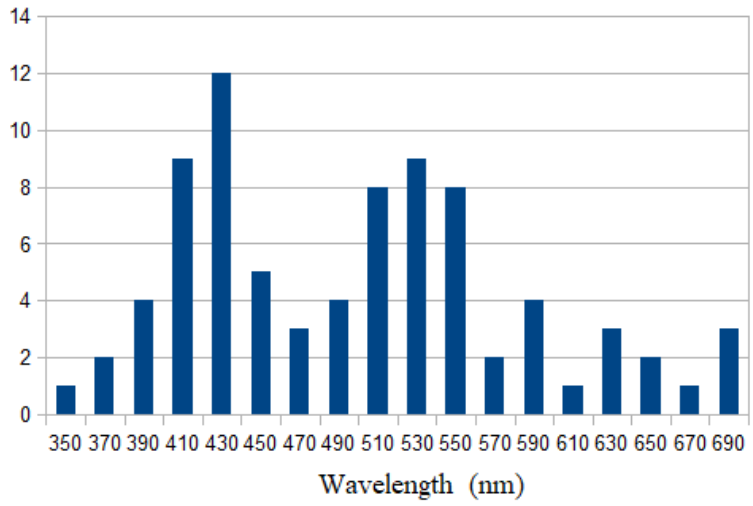

Figure 7: frequency histogram of the selections performed by SELECT

It is worthwhile to notice that the variables chosen by SELECT corresponded to the two bands highlighted by PARAFAC in the second mode, namely the emission spectra of two fluorophores. These outcomes are also in agreement with the profile of the loading on PC1, that highlights the presence of two important bands, the first positive at 410–450 nm and the second negative over 500 nm. Combining this information, it was possible to assume that the first emission band (410–450 nm) is due to a fluorophore characterizing the Chinese GT samples and that the broad band at 500–550 nm is related to the presence of compounds most abundant in the Japanese GT samples. The band at 410–450 nm probably corresponds to fluorescence emission of catechins, which are more abundant in Chinese samples. The band at 500–550 nm is probably attributable to carotenoids, that are recognized to be in particularly high quantities in Japanese tea, especially in Matcha, which contains 4 times more carotene than carrots and nine times more than spinach [27]. The infuses of GT prepared for the analysis were noticed to be slight yellow-green color due to pigments such as chlorophylls and carotenoids; the quantities of pigment extracted in hot water are related to the concentrations of the pigments in teas [28]. These observations agreed with the findings of Ref. [29], where the emission spectra of various organic compounds which are known to be endogenous component of plant leaves were measured, evidencing that catechins possess a fluorescence maximum near 440 nm and that β-carotene exhibits fluorescence emission with a maximum near 530 nm.

Conclusion and scientific impacts

The aim of the present study was to evaluate the possibility of using EEM fluorescence spectroscopy as a rapid analytical method for analyzing and characterizing GT samples, distinguishing between different geographical origins (China or Japan). The experimental data, given their complex and multivariate nature, were elaborated with chemometric techniques with the aim of extracting the useful information contained therein. PCA was applied, as a display technique, on the “unfolded data” and PARAFAC was performed on three-dimensional arrays. The PCA results were visualized by means of the score plot related to PC1 and PC4, which explained 76.8% of the total variance making it possible to distinguish Chinese and Japanese samples. The separation between the two geographical origins was mainly along PC1. Using PARAFAC, it was possible to perform the decomposition of the three-way excitation-emission matrix: the information on the first mode was similar to that observed by applying PCA to the matrix after unfolding and it demonstrated that fluorescence spectroscopy is a promising and fast analytical method to characterize GT samples on the basis of their geographical origin. PARAFAC on the second mode also highlighted the emission spectra of two fluorophores, one with a maximum around 420 nm and the other with a maximum at 500–550 nm. These bands correspond to the variables with the highest loadings on PC1 and also correspond to the variables selected by the SELECT algorithm, that are those with the highest discriminating power between Japanese and Chinese GT samples. The band around 420 nm was assumed to correspond to the fluorescence emission of catechins, which are more abundant in the Chinese samples, and the band around 500–550 nm was attributed to carotenoids. Moreover, the CyD-MEKC method was applied for the analysis of a subset of 24 GT samples confirming that catechins are more abundant in Chinese samples. In addition, the PLS-CM classification model built with these data made it possible to distinguish Japanese from Chinese GT samples with a sensitivity and specificity of 98.70% and 98.68%, respectively.

1.2 PROSTATE CANCER DETECTION BY EXCITATION-EMISSION FLUORESCENCE SPECTROSCOPY OF URIN COUPLED WITH CHEMOMETRICS

Scientific background and aim of the work

Prostate cancer is the second most widespread malignant tumour in the male population, after lung cancer, accounting for 1,276,106 new cases and causing 358,989 deaths in 2018 only. In Italy, it is the first cause of cancer death in males older than 60 years with more than 40,000 new cases per year and 18.5% of total cancers [30].

Early diagnosis of prostate cancer is often difficult in men over 60 years due to concomitant benign prostate hyperplasia which induces raised Prostate Specific Antigen (PSA) levels and improved prostate volume; the enlargement of the gland itself may not be indicative of the presence of a tumour but could also be due to a benign physiological proliferation.

The protocol for the prostate cancer diagnosis is preliminarily based on the urological examination, and the determination of PSA blood levels. Prostate biopsy is the only method able to confirm the presence of cancer cells in the prostate tissue, whilst multiparametric magnetic resonance investigation may be of some help for the non-invasive diagnosis of prostate cancer with a diagnostic accuracy of about 75%. Nevertheless, the lack of specificity of the PSA dosage and the prospect of preventing invasive and sometimes unnecessary prostate biopsies suggest the need of novel biomarkers or other non-invasive methods for the diagnosis of prostate cancer.

Based on the latest scientific evidence, it is of current interest to have rapid and accurate analytical methods for early screening of prostate cancer directly by urine analysis, in order to provide reliable results while improving patient compliance. For this purpose, non-destructive fingerprint spectroscopic methods, such as Fluorescence Spectroscopy, have proved to be extremely efficient for the analysis of biological fluids including urine, blood or plasma, in the clinical context [31].

In 2010, Masilamani et al. [32] showed the results of a novel study in which the native or intrinsic fluorescence of urine was used to aid the diagnosis of several types of cancer. In their study fluorescence emission spectra and Stokes shift spectra of the first voided urine samples were acquired for 100 healthy controls and those of 50 cancer patients of different aetiology. They concluded that flavoproteins and porphyrins released into urine can act as generic biomarkers of cancer with a specificity of 92%, a sensitivity of 76%, and an overall accuracy of 86.7%. A weak point of that study was that authors performed the statistical analysis to discriminate diseased patients from healthy patients using only seven a priori selected descriptors (ratios of intensity peaks) without considering, and benefiting from, the whole spectral information embodied into the fluorescence spectra.

In 2013, a study conducted by Zvarik et al. [33] dealt with assessing the differences in terms of metabolites, which can be detected by excitation-emission fluorescence, in urine samples of patients affected by ovarian cancer compared to healthy volunteers. They observed changes in the spectral profiles that were interpreted as reduction of pyridoxic acid content, whereas blue-fluorescing pteridines became dominant in urine samples of cancer patients with respect to healthy donors. Thus pteridines, which are related to cellular metabolism, could be suitable candidates for neoplasia-associated fluorescent markers in human urine. The observed changes in intrinsic fluorescence were studied by plotting as images (intensity represented by colour coding) the three-dimensional fluorescence excitation-emission landscapes, where the characteristic circular patterns, highlighting high emission, were used to identify wavelength regions which could be attributes to specific fluorophores. Also in this case, a chemometrics approach to extract the information from the whole excitation-emission landscapes for all samples (i.e. the raw measured complex analytical data) was not attempted.

In the same year, Rajasekaran et al. [34], studied native fluorescence characteristics of human urine samples using excitation–emission matrices (EEMs) over a range of excitation and emission wavelengths in order to discriminate patients with cancer from normal subjects. A total of 80 urine samples from normal subjects and 90 from pathologically confirmed cancerous patients were collected and analysed. EEMs were acquired in the following ranges, 250–450 nm for excitation and 270–750 nm for emission. However, only the spectra corresponding to fluorescence emission at 405

nm have been considered both for visual spectral comparison and to perform the discriminant analysis of normal form cancerous subjects. In more detail, stepwise multiple linear discriminant analysis was performed by the authors considering 19 ratio variables calculated using fluorescence intensities at emission wavelengths which represent characteristic spectral features of different groups of subjects, at 405 nm excitation.

The aim of the present study was to investigate excitation-emission Fluorescence Spectroscopy, as a rapid and accurate analytical method for the early screening of prostate cancer directly through urine analysis in order to provide reliable results while improving patient compliance.

An element of novelty of this study consists in processing the whole EEM landscapes with a suitable multiway analysis method, i.e., parallel factor analysis (PARAFAC) which allows resolution of the spectral profiles corresponding to single fluorophores and their relative concentration estimation. Indeed, the resolved profiles could be associated to chemical compounds (metabolites) that were recognised to have a role in oncological pathologies in literature. Thus, could be suggested as potential markers of prostatic oncological pathologies and subject to further investigation.

Experimental Plan: Sampling and Spectroscopic analysis

69 urine samples, provided by the Center of Urology University Hospital Cisanello of Pisa (Italy), were analysed; the samples belong to the following categories:

- 46 samples of patients whose prostate biopsy and subsequent histological examination have diagnosed a malignant prostate cancer.

- 23 control samples (healthy donors).

Urine samples were taken with a sterile procedure after pathological diagnosis of prostate cancer and before any kind of medical and/or surgical treatment. Urine samples were immediately frozen at - 80°C until analysis and stored according to hospital protocol; then, the samples were transported to the analytical laboratory in homologated packaging with inside dry ice pellets at -80 ºC and they were stored in a special cold room at -80 ºC.

Before being analysed, the samples were kept for 17 hours in the refrigerator (3 to 5 ºC) and for 1 hour in the thermostat rooms of the laboratory at a temperature of 20 ºC; since the samples in many cases had sediment, they were centrifuged for 30 minutes at 3000 rpm. Some urine samples were analysed in replicate to assess the repeatability of the method, and therefore the final number of EEM spectra that have been processed were 68 from cancer patients and 29 from healthy donors.

The EEM fluorescence measurements were performed on centrifugated urine samples at room temperature on a Perkin-Elmer LS55B luminescence spectrometer (Waltham, MA, USA). The excitation- emission matrices (EEMs) of the urine were recorded using the standard cell holder in a 10 mm quartz SUPRASIL® cell with cell volume of 3.5 mL by PerkinElmer. The excitation spectra were recorded between 250 nm and 530 nm each 5 nm (29 recorded points), whereas the emission wavelengths ranged from 270 nm to 650 nm each 0.5 nm (761 points). The excitation and the emission monochromator slits were set to 10 nm. The FL WinLab software (PerkinElmer) was used to register the fluorescent signals.

Chemometric approach purposed

The acquired EEM landscapes were arranged in a three-way data array of dimensions 97 (human urine samples) x 761 (emission wavelengths) x 29 (excitation wavelengths). The data were split into a calibration (70%, i.e. 69 samples) and a validation (30%, i.e. 28 samples) sets, by using the Kennard Stone Duplex algorithm [6] distinct per category in order to keep the same percentage of calibration and validation samples for each class (proportional splitting).

Parallel Factor Analysis (PARAFAC)

According to the specific nature of EEM data, organised in a three-way data array (sample × λ emission × λ excitation), Parallel Factor Analysis algorithm [3], was applied to model directly the n-way data [5].

$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$ i = 1,2, ..., I; j = 1,2, ..., J; k = 1,2, ..., K

EEM data were first pre-processed in order to minimise the non-relevant instrumental artefacts. In particular, Rayleigh scatter, normally present in this kind of data [10], was removed using a first and a second order Rayleigh filters (half-width: 20 nm) and replaced with interpolated data. Zeros were assigned to sub-Rayleigh wavelengths [5]. Then, filtering in the second mode (eleven points window) and despiking were applied. In order to select the proper number of PARAFAC factors, different parameters have been considered: the total variance explained by the model (fit), the core consistency; the similarity of fit and core consistency for replicate runs, i.e. PARAFAC has been restarted 5 times for each model dimensionality explored (from 1 to 6); the congruence of mode 2 and mode 3 loadings profiles obtained by split half analysis, i.e. obtained by dividing the data set in two parts with respect to samples mode and calculating a distinct PARAFAC model on each sub-set. The congruence is estimated as the covariance between corresponding loadings of the two halves. These criteria are implemented in the PLS toolbox, namely *nvalidate* function.

Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) [35] was carried out by using the scores of PARAFAC model as class descriptors.

Partial Least Square Discriminant Analysis

Partial least squares discriminant analysis (PLS-DA) [36], was applied in order to discriminate cancer patients from healthy donors using as independent variables the first mode scores (relative concentrations) of PARAFAC model. The number of PLS-

DA components was chosen according to minimum root mean squares error in Leave One Out cross validation.

Data were imported to MATLAB v. R2019a (The MathWorks, Inc., Natick, MA, US). PARAFAC was performed using PLS Toolbox v. 8.9 (Eigenvector Research, Inc., Manson, WA, US). Linear discriminant analysis was performed by using the functions of the Statistical and Machine Learning Toolbox of MATLAB.

Research outcomes

The analysis of EEM data shows that interesting differences in urine fluorescence excitation/emission spectra from patients with prostate cancer in comparison to healthy subjects can be already appreciated starting from the collected data, after a pre-processing step applied to remove the interference from Rayleigh scattering. As an example, two of the fluorescence excitation-emission landscapes of urine matrices coming from a healthy person (H16 sample) and a prostate cancer patient (P03) are presented in Figure 1a and 1b, as contour plots.

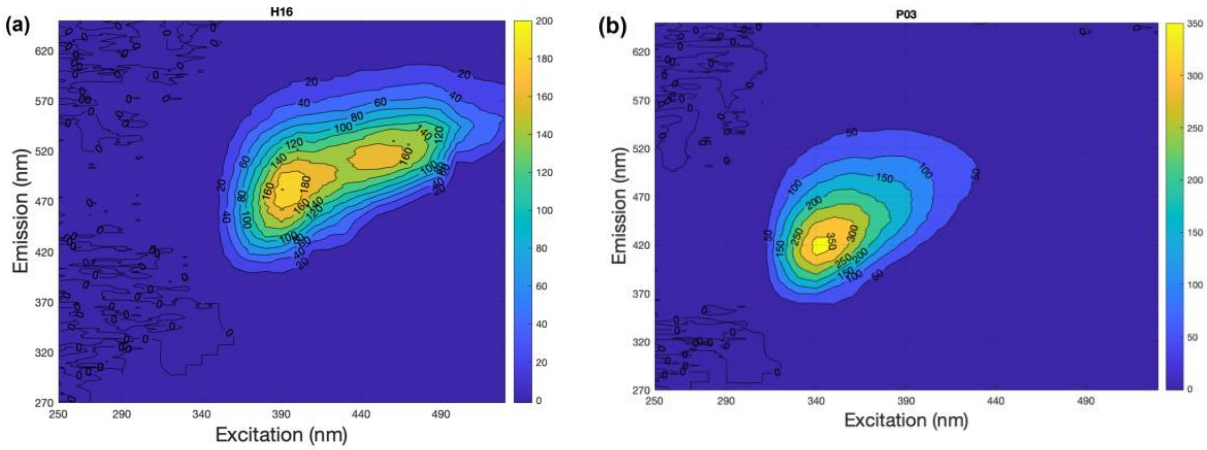

Figure 1: (a) contour plot related to a patient affected by prostate cancer. (b) Contour plot related to a healthy subject

From Figure 1, it is quite evident the presence of strongly overlapping fluorescence bands in both EEM spectra; however, some differences in the landscape of the two samples could be highlighted. Firstly, in the healthy spectrum, two main bands could

be visualized at 390 (excitation)/490 (emission) nm and at 470/510 nm with a shoulder at 490/520. These characteristics were observed in almost all the urine samples coming from healthy controls. Conversely, there is a fading of the peak at 490/520 nm in cancer urine sample, coupled with the emergence of the band at 350/420 nm. Furthermore, the reported cancer urine spectrum is characterized by a depression of fluorescence at 390/490 nm with respect to the one of healthy sample. However, not all the urine spectra from the 46 cancer patients showed the same trend; in general, among cancer patients, two different behaviours seem to be present: one characterized by one main high intensity band at around 350/420nm and another characterized by a low intensity of the whole excitation/emission landscape. However, it is very hard to decipher the different emission bands due to the complexity of the matrix and the presence of overlapping bands. Thus, a multiway resolution method, i.e. PARAFAC, has been applied, benefiting from the second order advantage, in this way a clearer interpretation of the highlighted bands could be achieved.

PARAFAC analysis

The EEM data were arranged in a three-way data array of dimensions I x J x K, where I is the number of investigated samples (97 human urines samples in total), J the number of emission wavelengths (761 points) and K the number of excitation wavelengths (29 points). Before decomposition by PARAFAC, the data were appropriately pre-treated in order to minimize the non-relevant instrumental artifacts . A four-factors PARAFAC model was selected in according to criteria based on residuals, core consistency and split-half analysis [5], indeed four different fluorophores were detected in the investigated urine samples.

The results from the PARAFAC model are reported in Figures 2a and 2b. In particular, Figure 2a shows the emission (mode 2 loadings) and Figure 2b the excitation (mode 3 loadings) profiles of the four resolved factors in urine samples. These two modes represent the underlying pure spectra of characteristic fluorophores present in the investigated urine samples and which can be putatively identified based on literature.

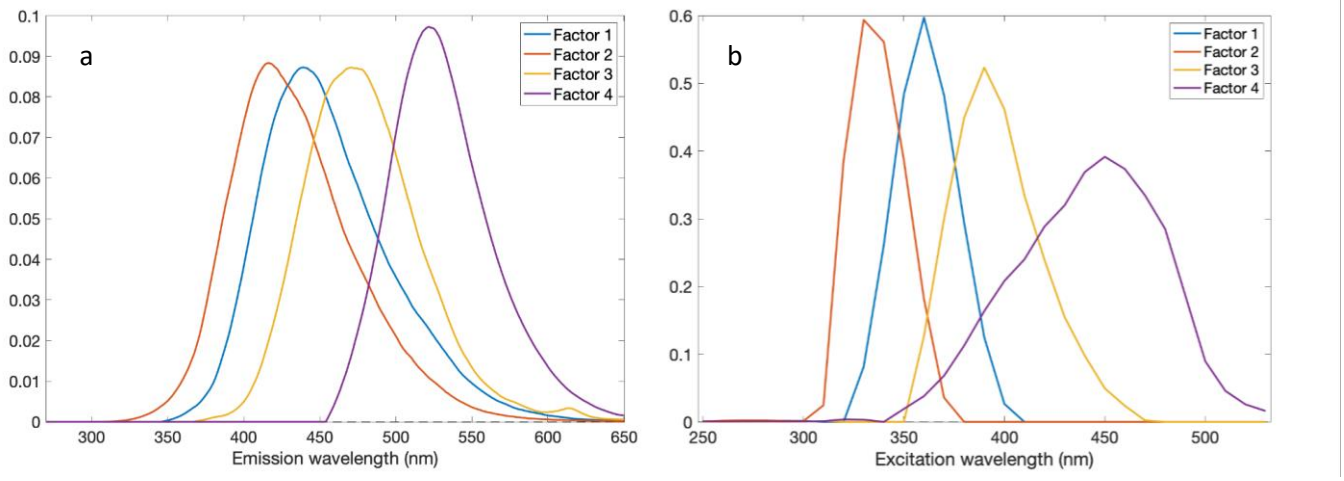

Figure 2: (a) Emission loading vectors from four factors PARAFAC model. (b) Excitation loading vectors from four factors PARAFAC model

The excitation/emission wavelengths corresponding to the maximum fluorescent intensity of the first factor (blue) is 360/440 nm. According to several authors, the band at around 370/440 nm could be ascribed to one or more chemical species, such as pteridines [33]. These compounds could play important roles in the synthesis of some vitamins, as well as they are important intermediates in anabolic and catabolic reactions. Furthermore, Masilamani et al. [32] reported that the 444 nm band in emission could be due to NADH bound to a protein, even if it could be difficult to envisage the presence of this molecule in healthy urine samples [33]. Although shifts are present with respect to the reference for both NADH and pteridines, they could be acceptable since it is well known that the fluorescent emission signal from a fluorophore can be strongly dependent on the surrounding environment [37].

The second (red) factor has excitation and emission maxima at 330 and 420 nm, respectively, and could be attributed to several fluorophores such as pyridoxic acid and uric acid [33]. In particular, pyridoxic acid is excreted in the urine as a catabolic product of vitamin B6 and it is involved in many enzymatic reactions as pyridoxal-phosphate active form of vitamin B6. The excitation/emission maxima positioned at 390/470 nm characterizes the third factor. Free NADH has an excitation/emission maximum at around 400/485 nm [32], hence, considering a potential shift of the

emission when the polarity of the microenvironment changes, this band could be tentatively associated with the third PARAFAC factor. Furthermore, from a closer inspection of the shape of this factor, it is worth noticing a little, hardly visible ‘‘bump’’ with emission maximum at 620 which is in agreement with literature values for porphyrins [32].

Finally, excitation/emission loadings (450/530 nm) of the fourth factor could fit well with excitation/emission of flavins and their metabolites [33]. In particular, riboflavin presents an excitation/emission maximum at around 450/550 nm [38]. Furthermore, Masilamani et al. [32] reported that bilirubin could also contribute to the fluorescence band at these wavelengths and its band could overlap in this region. However, flavin is more fluorescent than bilirubin, therefore, the latter could not apport a different and prominent contribution. The role of the four PARAFAC factors to distinguish healthy and prostate cancer samples can be inspected by first loading mode plots shown in Figure 3a and Figure 3b.

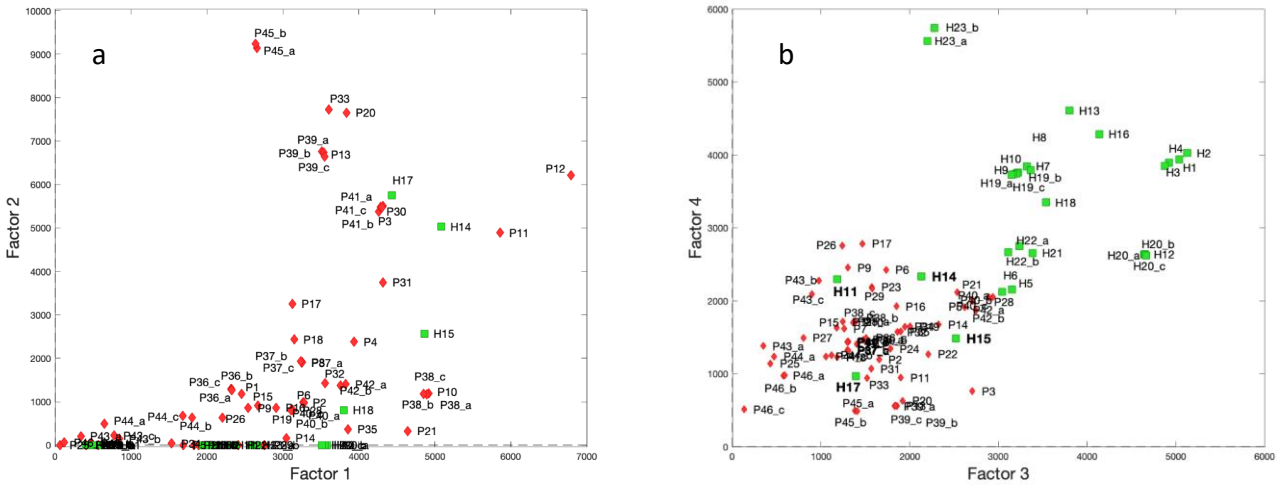

Figure 3: Mode 1 PARAFAC loadings plot of (a) first vs. second PARAFAC factors and (b) third vs. fourth ones.

Figure 3a displays the first PARAFAC loading mode plot for the first vs. the second factors: first, a good reproducibility within the different replicates (same number in the label) is observed along both the factors, on the other hand the samples spread in the scores is not related to a clear cancer/non cancer distinction. Rather a differentiation into two groups arises: one heterogeneous, includes almost all cancer samples that, together with H14, H15, H17 and H18 healthy controls, and lies at high score values for both factors, i.e., pteridines and pyridoxic acid; the second one, is formed by several samples positioned at factor 2 (pyridoxic acid) scores value, close to zero. Pteridines have been already shown to be an interesting neoplasia marker in human urine since their biosynthesis could be altered by the presence of malignant tumors which lead to a change of their concentration [33, 39]. As regards pyridoxic acid, it is one of the catabolic products of vitamin B6 thus, its higher amount in cancer samples could reflect a decrease of the vitamin B6, whose concentration has been hypothesized to reduce cancer risk [40]. Although no prior information about the healthy status of the patients has been used for building the PARAFAC model (i.e., it is an unsupervised method), it was possible to highlight a partial separation between healthy and diseased individuals (Figure 3b) considering the third and fourth factors. In particular, it is worth noting an increase of both the third (putative free form of NADH) and the fourth (flavins) factors passing from urine samples of cancer patients to healthy donors. Four samples (H11, H14, H15 and H17) of healthy donors, nonetheless present an ‘anomalous’ trend and further investigation are ongoing to better understand their behaviour.

LDA and PLS-DA results

The PARAFAC results discussed above showed that factors 3 and 4 are the most significative in differentiating between healthy volunteers and prostate cancer patients. Thus, linear discriminant analysis was applied to factors 3 and 4 to obtain a classification model. The results are shown in Figure 4.

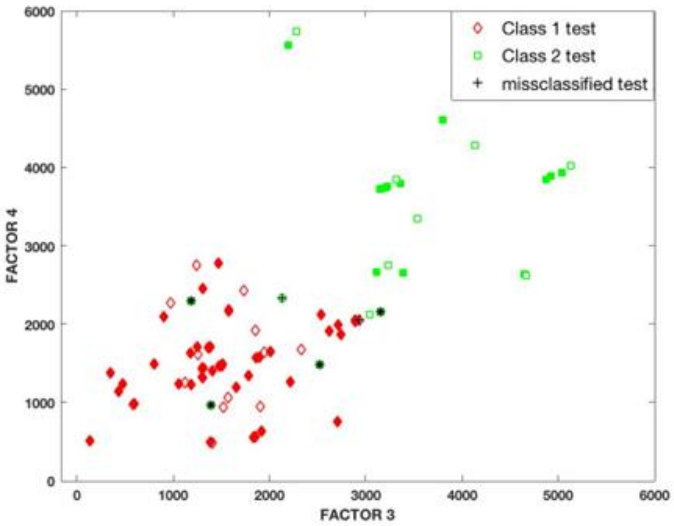

Figure 4: Plot of mode 1 PARAFAC loadings: factor 3 vs. factor 4; filled signs indicate calibration set and empty markers validation set. The asterisks and crosses indicate samples misclassified by LDA for calibration and validation set, respectively.

As concern the calibration set, four healthy samples are misclassified, three of these, H11, H15 and H17, have been already noticed in the exploratory analysis and discussed above; on the other hand, all the prostate cancer samples are correctly classified. For the test set, one healthy (H14, the fourth previously discussed) and one cancer (M33) are respectively misclassified. The overall performance of the LDA model is reported in Table 1.

Table 1: LDA classification results

	Calibration set	Validation set	Sensitivity training	Specificity training	Sensitivity test	Specificity test
Cancer class	47	19	100%	80%	94,7%	88,9%
Healthy class	20	9	80%	100%	88,9%	94,7%

These results are extremely encouraging, since only one cancer sample in prediction was misclassified, and the same for healthy ones. The four healthy samples as already discussed show rather different EEM landscapes with respect to the others and should be further investigated. For comparative purposes, a PLS-DA analysis has been also performed by considering all the four PARAFAC factors. The model was cross-validated using a venetian blind procedure with five splits, two latent variables were selected and the results in terms of sensitivity and specificity are reported in Table 2.

Table 2: Results from PLS-DA model for classification of cancer and healthy classes based on PARAFAC scores

	Calibration set	Validation set	LVs	AUC (CV)	Sensitivity CV	Specificity CV	Sensitivity predict	Specificity predict
Cancer class	47	19	2	0,92	91,5%	85%	94,7%	88,9%
Healthy class	20	9			85%	91,5%	88,9%	94,7%

The area under the ROC curve (AUC in CV) is rather high (Table 2), however three of the healthy samples were wrongly assigned to cancer class and four of the latter were incorrectly predicted as healthy ones, both in fit and cross-validation. It has to be noticed that the three healthy samples accepted by the cancer class are the same H11, H15 and H17, misclassified by the linear discriminant model. The prediction capability improved in validation where only one sample per class was misclassified; finally, the healthy misclassified sample is H14 and the cancer one is the same one misclassified by LDA (M33). More trustworthy results could be surely obtained by enlarging the number of samples, anyhow the results are very consistent.

A biplot for the PLSDA model of cancer vs. healthy findings based on the PARAFAC scores is reported in Figure 5. As expected from the sensitivity and specificity values, there is an almost good separation between the two classes (except for four healthy samples).

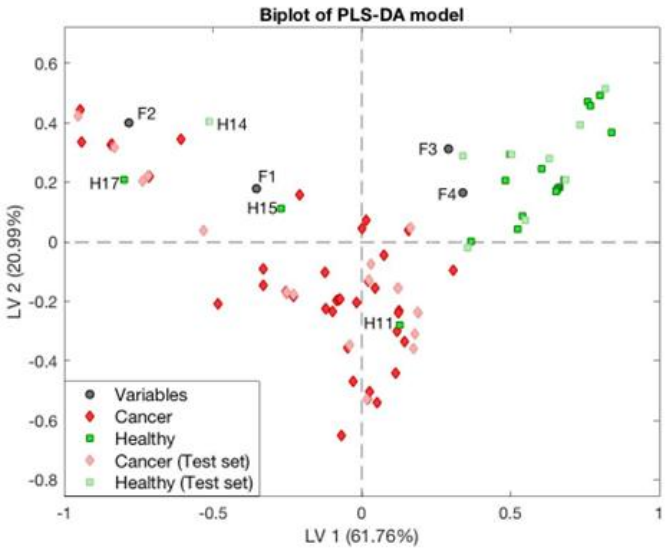

Figure 5: PLS-DA biplot of the first vs. second PLS-DA components. Cancer patients (red diamonds); healthy persons (green circles). Variables, i.e. mode 1 PARAFAC loadings (gray circles), F1, F2, F3 and F4 stand for first, second, third and fourth components.

In particular, there is a tendency towards differentiation along the first PLS-DA component, whilst on the second it is possible to note a difference in the intra cancer class which is split into two groups, the first one with positive second component scores and the second group with negative ones. In the biplot, it is also possible to see which variables are important for this separation. In particular, the samples that are positively correlated to the ‘‘cancer direction’’ present a lower amount for both 3 and 4 variables (i.e. free-NADH and flavins) with respect to the control samples. Finally, variables 1 and 2, pteridines and pyridoxic acid respectively, contribute to differentiate the two groups among cancer samples, showing that almost all cancer samples, with both LVs negative scores, generally present a lower amount of the intensity bands for all the four fluorophores.

Conclusion and scientific impacts

In this study, Excitation-Emission fluorescence measurements on human urine combined with multivariate data analysis were proposed as a potential fast routine analysis to screen potential prostate cancer patients. Previous studies with fluorescence spectroscopy have shown optimal results of EEM in the investigation of cancer patients of different aetiology, but this work, to the best of author's knowledge, represents the first attempt in the discrimination of prostate cancer patients.

On interpretative ground, four fluorophores, corresponding to the four PARAFAC factors, were resolved and could act as potential markers in the differentiation between urine samples from healthy donors and cancer patients By a comparison with the results of previous studies performed on urine, these could be putatively assigned as: pteridines and/or bounded NADH at 360/460 nm (excitation/emission); pyridoxic acid at 330/420 nm; and, free-NADH and flavins, in the regions at 390/470 nm and 450/530 nm, respectively. The first two compounds seem to be characteristic only of a subgroup of cancer samples with higher concentrations, while the latter two contributed to some extent in the differentiation of healthy from cancer samples, which present lower values of concentration of both the fluorophores.

PARAFAC allowed enhanced interpretation of the results, thanks to its capability to furnish resolved factors chemically interpretable. The analysis can be further improved by using standard addition of putative analytes (species) identified in the characterization phase. Moreover, PARAFAC scores, which represent the relative concentrations of the resolved species, when fed to LDA and to PLS-DA model, furnished a first evaluation of the capability to achieve healthy/cancer discrimination. The relatively low number of samples prevent to be conclusive about the robustness of the obtained results. However, the obtained results could serve as a first basis to the development of a simple and non-invasive protocol for prostate cancer detection to be used as a screening tool able to support the different techniques used in this issue.

[1] Lakowicz, J. R. (Ed.). (2006). Principles of fluorescence spectroscopy. Boston, MA: springer US.

[2] Kasha, M. (1950). Characterization of electronic transitions in complex molecules. Discussions of the Faraday society, 9, 14-19.

[3] Bro, R. (1997). PARAFAC. Tutorial and applications. Chemometrics and intelligent laboratory systems, 38(2), 149-171.

[4] Lenhardt, L., Bro, R., Zeković, I., Dramićanin, T., & Dramićanin, M. D. (2015). Fluorescence spectroscopy coupled with PARAFAC and PLS DA for characterization and classification of honey. Food Chemistry, 175, 284-291.

[5] Bro, R., & Kiers, H. A. (2003). A new efficient method for determining the number of components in PARAFAC models. Journal of Chemometrics: A Journal of the Chemometrics Society, 17(5), 274-286.

[6] Guimet, F., Ferré, J., Boqué, R., & Rius, F. X. (2004). Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation–emission matrix fluorescence spectroscopy. Analytica Chimica Acta, 515(1), 75-85.

[7] Ortiz, M. C., Sarabia, L. A., & Sánchez, M. S. (2010). Tutorial on evaluation of type I and type II errors in chemical analyses: From the analytical detection to authentication of products and process control. Analytica Chimica Acta, 674(2), 123-142.

[8] Suzuki, Y., Miyoshi, N., & Isemura, M. (2012). Health-promoting effects of green tea. Proceedings of the Japan Academy, Series B, 88(3), 88-101.

[9] Bogdanski, P., Suliburska, J., Szulinska, M., Stepien, M., Pupek-Musialik, D., & Jablecka, A. (2012). Green tea extract reduces blood pressure, inflammatory biomarkers, and oxidative stress and improves parameters associated with insulin resistance in obese, hypertensive patients. Nutrition research, 32(6), 421-427.

[10] Cabrera, C., Artacho, R., & Giménez, R. (2006). Beneficial effects of green tea—a review. Journal of the American College of Nutrition, 25(2), 79-99.

[11] Cooper, R. (2012). Green tea and theanine: health benefits. International journal of food sciences and nutrition, 63(sup1), 90-97.

[12] Velayutham, P., Babu, A., & Liu, D. (2008). Green tea catechins and cardiovascular health: an update. Current medicinal chemistry, 15(18), 1840.

[13] Wang, H., Provan, G. J., & Helliwell, K. (2000). Tea flavonoids: their functions, utilisation and analysis. Trends in Food Science & Technology, 11(4-5), 152-160.

[14] Yuan, J. M., Sun, C., & Butler, L. M. (2011). Tea and cancer prevention: epidemiological studies. Pharmacological research, 64(2), 123-135.

[15] Bonoli, M., Colabufalo, P., Pelillo, M., Gallina Toschi, T., & Lercker, G. (2003). Fast determination of catechins and xanthines in tea beverages by micellar electrokinetic chromatography. Journal of agricultural and food chemistry, 51(5), 1141-1147.

[16] Kosińska, A., & Andlauer, W. (2014). Antioxidant capacity of tea: effect of processing and storage. In Processing and impact on antioxidants in beverages (pp. 109-120). Academic Press.

[17] Pasquini, B., Orlandini, S., Goodarzi, M., Caprini, C., Gotti, R., & Furlanetto, S. (2016). Chiral cyclodextrin-modified micellar electrokinetic chromatography and chemometric techniques for green tea samples origin discrimination. Talanta, 150, 7-13.

[18] Ma, G., Zhang, Y., Zhang, J., Wang, G., Chen, L., Zhang, M., ... & Lu, C. (2016). Determining the geographical origin of Chinese green tea by linear discriminant analysis of trace metals and rare earth elements: Taking Dongting Biluochun as an example. Food Control, 59, 714-720.

[19] Diniz, P. H. G. D., Barbosa, M. F., de Melo Milanez, K. D. T., Pistonesi, M. F., & de Araújo, M. C. U. (2016). Using UV–Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup. Food chemistry, 192, 374-379.

[20] Ye, N. S. (2012). A minireview of analytical methods for the geographical origin analysis of teas (Camellia sinensis). Critical reviews in food science and nutrition, 52(9), 775-780.

[21] Ye, N., Zhang, L., & Gu, X. (2012). Discrimination of green teas from different geographical origins by using HS-SPME/GC–MS and pattern recognition methods. Food Analytical Methods, 5(4), 856-860.

[22] Gotti, R., Furlanetto, S., Lanteri, S., Olmo, S., Ragaini, A., & Cavrini, V. (2009). Differentiation of green tea samples by chiral CD‐MEKC analysis of catechins content. Electrophoresis, 30(16), 2922-2930.

[23] Ortiz, M. C., Sarabia, L. A., Sánchez, M. S., & Giménez, D. (2009). Identification and quantification of ciprofloxacin in urine through excitation-emission fluorescence and three-way PARAFAC calibration. Analytica chimica acta, 642(1-2), 193-205.

[24] Forina, M., Lanteri, S., Casale, M., & Oliveros, M. C. C. (2007). Stepwise orthogonalization of predictors in classification and regression techniques: An “old” technique revisited. Chemometrics and Intelligent Laboratory Systems, 87(2), 252-261.

[25] Forina, M., Lanteri, S., Armanino, C., Casolino, C., Casale, M., & Oliveri, P. (2003). V-PARVUS. An extendable package of programs for explorative data analysis, classification and regression analysis. Dip. Chimica e Tecnologie Farmaceutiche ed Alimentari, University of Genova.

[26] MATLAB, V. (2014). 8.4. 0.150421 (R2014b). The MathWorks Inc., Natick, Massachusetts.

[27] Hall, N. (2000). The tea industry. Elsevier.

[28] Suzuki, Y., & Shioi, Y. (2003). Identification of chlorophylls and carotenoids in major teas by high-performance liquid chromatography with photodiode array detection. Journal of Agricultural and Food Chemistry, 51(18), 5307-5314.

[29] Lang, M., Stober, F., & Lichtenthaler, H. K. (1991). Fluorescence emission spectra of plant leaves and plant constituents. Radiation and environmental biophysics, 30(4), 333-347.

[30] Spandonaro, F., D’angela, D., Polistena, B., Bruzzi, P., Iacovelli, R., Luccarini, I., ... & Brigido, A. (2021). Prevalence of Prostate Cancer at Different Clinical Stages in Italy: Estimated Burden of Disease Based on a Modelling Study. Biology, 10(3), 210.

[31] Madhuri, S., Vengadesan, N., Aruna, P., Koteeswaran, D., Venkatesan, P., & Ganesan, S. (2003). Native Fluorescence Spectroscopy of Blood Plasma in the Characterization of Oral Malignancy¶. Photochemistry and Photobiology, 78(2), 197-204.

[32] Masilamani, V., Trinka, V., Al Salhi, M., Govindaraj, K., Raghavan, A. P. V., & Antonisamy, B. (2010). Cancer detection by native fluorescence of urine. Journal of biomedical optics, 15(5), 057003.

[33] Zvarik, M., Martinicky, D., Hunakova, L., Lajdova, I., & Sikurova, L. (2013). Fluorescence characteristics of human urine from normal individuals and ovarian cancer patients. Neoplasma, 60(5), 533-537.

[34] Rajasekaran, R., Aruna, P. R., Koteeswaran, D., Padmanabhan, L., Muthuvelu, K., Rai, R. R., ... & Ganesan, S. (2013). Characterization and diagnosis of cancer by native fluorescence spectroscopy of human urine. Photochemistry and photobiology, 89(2), 483-491.

[35] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.

[36] Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. Journal of Chemometrics: A Journal of the Chemometrics Society, 17(3), 166-173.

[37] Abugo, O. O., Nair, R., & Lakowicz, J. R. (2000). Fluorescence properties of rhodamine 800 in whole blood and plasma. Analytical biochemistry, 279(2), 142-150.

[38] Lenhardt, L., Bro, R., Zeković, I., Dramićanin, T., & Dramićanin, M. D. (2015). Fluorescence spectroscopy coupled with PARAFAC and PLS DA for characterization and classification of honey. Food Chemistry, 175, 284-291.

[39] Gamagedara, S., Gibbons, S., & Ma, Y. (2011). Investigation of urinary pteridine levels as potential biomarkers for noninvasive diagnosis of cancer. Clinica Chimica Acta, 412(1-2), 120-128

[40] Larsson, S. C., Orsini, N., & Wolk, A. (2010). Vitamin B6 and risk of colorectal cancer: a meta-analysis of prospective studies. Jama, 303(11), 1077-1083.

CHAPTER 2: NEAR INFRARED SPECTROSCOPY, BENCHTOP APPLICATIONS.

The discovery of near-infrared radiation was ascribed to William Herschel, astronomer and scientist, in the 1800s [1]. Only much later, around 1950s, the Near Infrared Spectroscopy (NIRs) was applied for industrial applications. A scientist named Karl Norris, who was working for US Department of Agriculture tested for a first time, an innovative analytical approach based on NIRs for analysing grain samples in a fast and non-destructive way. This delay in the use of this technique was because of the difficulty in extracting the information from the broad and overlapped bands that characterize NIR spectra. Thanks to the introduction of the first single-unit, stand-alone NIRS system in 1980s, light-fiber optics in mid1980s, and the development of monochromator detector in 1990s, NIRS became a more powerful tool for scientific research [2]. Moreover, the progresses in terms of computational power and the wider application of multivariate data analysis have proven the advantages in using NIRS in different fields as pharmaceutical [3], agriculture [4], chemical [5] and food industry [6].

The versatility of this technique is related to the possibility of acquiring, in few seconds and without sampling preparation, a lot of interesting chemical and physical information about a sample of interest.

When the absorption of the NIR radiation is equal to the difference between two vibrational energy levels, the molecules of the sample interact with the frequencies of the light. Some frequencies of the incident light are absorbed by the sample, while the others can be partially absorbed or not absorbed at all. In general, the absorption takes place only if the vibrational movement of the atoms that form the molecular bond or the atoms forming a local group of vibrating atoms, creates a change in the dipole moment.

For this reason, absorptions in the NIR region (780–2526 nm) are primarily due to overtones and combination bands of the fundamental molecular vibrations that occur when transitions to excited states involve two vibrational modes at the same time.

Overtone absorption bands originate from the functional groups that contain C-H, N-H, O-H or S-H atomic bonds. Overtone vibrations that originate from the covalent bonds are combined with lower-frequency fundamental bands such as C=O and C-C to generate overtone-combination bands. The intensity of the absorption bands depends on the degree of the dipole change during the vibration of the bonds. In according to the nature of the sample, NIR spectra can be obtained in three different modes: transmission, diffuse reflection or transflection. In this work, benchtop instruments working both in transmission and in diffuse reflectance mode have been used for acquiring respectively liquid and solid samples. In diffuse reflectance spectroscopy, the light source and detector are located on the same side of the sample. The detector measures the radiation reflected from the sample surface, which contains a specular component and a diffuse component. Transmission mode is usually applied for analysing transparent liquid samples. In this case the light source is located on the opposite side from the sensor which records the light transmitted through the sample.

Based on the wavelength selection, it's possible to classify the modern NIR instrumentation in 4 categories:

1) Filter based instruments

2) LED based instruments

3) Acousto-Optical Tunable Filters (AOTF) based instruments

4) Fourier-transform (FT) based instruments

In this second chapter, two different applications in which a FT-NIR benchtop spectrophotometer has been used to perform the analysis have been proposed. In respect to the other categories of instruments, FT-NIR spectrometers offer several advantages in terms of wavelength precision and accuracy, high signal to-noise ratio, scan speed and versatility in the sample presentation.

The first paragraph of this second chapter reports a food application in which a Buchi NIRFlex N-500 benchtop spectrophotometer has been used for analysing extra virgin olive oil (EVOO) samples. Nowadays, laboratories that perform 'highly frequent' analysis on EVOO by NIRS usually employ quartz cuvettes. This results in time-

consuming measurements, especially in the cleaning phase, and an increased cost for non-green cleaning solvents. The use of disposable glass vials may reduce time and costs significantly, but their analytical performances in EVOO analysis, have not yet been investigated. In order to reach this goal, a set of 106 EVOO samples from different Italian olive-growing areas were collected and analysed using both quartz cuvette and mono-use glass vials. The analytical performances of the cuvettes have been tested in terms of multivariate calibration models were developed to estimate quality parameters of extra virgin olive oil. The quantitative models have been developed by the application of Partial Least Square (PLS) regression algorithm [7]. PLS works maximizing the covariance between the original data matrix $\mathbf{X}$ (NIR spectra) and the matrix of reference parameters to be predicted $\mathbf{Y}$. This covariance information is expressed summarized in few successive abstract factors, called latent variables. The PLS algorithm decomposes the matrices X and Y in factor scores T and U related to samples included in X and Y, respectively, and factor loadings P and Q related to variables in X and Y, respectively. The factor decomposition can be expressed by the equations below:

$$\mathrm{X} = \mathrm{TP}^{\mathrm{T}} + \mathrm{E}$$

$$\mathrm{Y} = \mathrm{UQ}^{\mathrm{T}} + \mathrm{F}$$

where E and F are the residuals in X and Y, respectively, that corresponds to the information not taken into account by the model. The regression model is obtained by the eq. using T and U.

$$\mathrm{U} = \mathrm{TB}_{PLS} + \mathrm{H}$$

where B_{PLS} is a matrix that contains the PLS coefficients relating the information of X and Y, expressed by their respective scores.

The predictive ability of each PLS model was evaluated on an independent test set. The Passing-Bablok linear regression [8] was lastly used to statistically compare the performances of the two different types of cuvettes.

The second study reported in this Chapter was focused in finding the relationship between the water activity and the water molecular structure of the rice germ, based on its spectral pattern which can be measured using non-destructive technology. Aquaphotomics [9] near-infrared spectroscopy was used to study rice germ stored at different levels of water activity and atmosphere. Aquaphotomics is a new “omics” discipline introduced by Professor Roumiana Tsenkova at the Laboratory of Bio Measurement Technology at Faculty of Agriculture, Kobe University, Japan. This approach is based on the key role of the water in biological and aqueous systems. The use of the NIR radiation allows to measure specific regions of the spectrum of the light that comes back out of the water getting in this way information about changes of water molecular vibrations in relation to other molecular vibrations present in the system. All 12 water absorbance bands called Water Matrix Coordinates (WAMACS) have been used to define the water absorption spectral pattern (WASP), which describes the condition of the whole aqueous system and allows the representation of the specific absorption pattern in dedicated radar plots, called aquagrams.

Chapter 2

2.1 A CHEMOMETRIC STRATEGY TO EVALUATE THE COMPARABILITY OF PLS MODELS OBTAINED FROM QUARTZ CUVETTES AND DISPOSABLE GLASS VIALS IN THE DETERMINATION OF EXTRA VIRGIN OLIVE OIL QUALITY PARAMETERS BY NIR SPECTROSCOPY.

Scientific Background and aim of the work

The International Olive Oil Council (IOOC) fixed purity and quality criteria in order to recognize four commercial olive oil categories (or grades): “extra-virgin” olive oil, “virgin” olive oil, “refined” olive oil and “pomace” oil [10]. Extra-virgin olive oil (EVOO) is considered the highest quality grade and the adulteration with edible oil of inferior quality is becoming a type of commercial fraud more and more frequent. The quality criteria established by the IOOC for EVOO include measurements related to sensory characteristics (odour, taste and colour), free acidity, peroxide value, absorbance in the ultra-violet spectral region at 232 and 270 nm (K 232, K 270, ΔK), moisture and volatile matter. In addition to these main physicochemical parameters, the content of methyl esters of fatty acids (FAMEs) and triacylglycerols (TAGs) represent important parameters for characterizing olive oil samples [11]. These compounds are considered particularly interesting for their physiological effects [12] and suitable for authenticity assessment of EVOO [13]. In this context, to ensure the highest quality of the Italian EVOO and to counter fraudulent trade, the Violin project (Valorisation of Italian Olive Products Through Innovative Analytical Tools), promoted by Ager foundation, designed innovative analytical protocols, including approaches based on near infrared spectroscopy (NIRS) and multivariate data analysis.

It is well known, in fact, that NIRS nowadays represents a valid and recognized alternative method, compared to traditional techniques, to determine qualitative and quantitative parameters of several food matrices, including olive oil, in a rapid and non-destructive way, requiring no or limited sample preparation, with a reduction of costs and time of analysis. In the literature, in fact, there are several studies that proved the

potential of NIRS technology for determining the quality of olive oil, both in terms of chemical composition [14] and product authentication [15]. Regarding chemical composition, NIRS has been demonstrated to be useful for quantifying important control parameters, including peroxide value, free fatty acid content, and specific extinction coefficients (e.g. K232 and K270) [16]. Regarding food frauds, NIRS has proven to be an effective analytical method to detect and estimate adulteration of olive oils with vegetable oils of inferior quality [17]. Moreover, in the last decades, NIR spectroscopy has been recognized as an excellent tool for the verification of authenticity of EVOO samples concerning geographical origin [18] or olive cultivar.

The main advantage of the NIR technique, is that it is a quick and low-cost method, but the speed of spectra acquisition can be limited by the employment of quartz cuvettes, especially due to the cleaning phase, which often includes the use of organic solvents, such as acetone, resulting in a non-green methodology. In addition, an improper use of these chemicals also can leave residues in cuvettes, leading to possible signal alterations.

The introduction on the market of disposable optical glass vials (DGV) may reduce acquisition time and costs both in industry and in research laboratories; these vials are much cheaper but can have slightly different geometries and the thickness of the glass differ between one vial and the other and even inside the same cuvette. Furthermore, due to differences between optical glass and quartz in terms of transmission range, thermal properties and chemical compatibility, a critical comparison between these two types of cuvettes is required and it has not yet been investigated, in particular for the analysis of olive oil. To this aim, the development of a suitable chemometric strategy is a fundamental step.

Given these premises, the present comparative study was aimed at understanding if the use of DGV for the NIRS analysis could significantly affect the prediction of quality parameters in EVOO samples. To reach this goal, a total of 106 EVOO samples were acquired with the same NIRS device, using both quartz cuvettes (QC) and DGV. On the spectra obtained, an optimization step of data pre-processing was carried out and, then, Partial Least Squares (PLS) regression [19] was applied on a calibration set of the NIRS data to develop quantitative models for FAMEs and TAGs content. The

prediction ability of these models was estimated on a separate test set. The values predicted by these models obtained from spectra recorded using QC and DGV were used to compare, for the first time, the analytical performances of these two types of cuvettes. To do this, the symmetrical and non-parametric Passing-Bablok regression method – applicable also when the x variable has substantial uncertainties – was applied on the regression models obtained; a joint statistical test on slope and intercept was performed considering the null hypothesis (H_0) verified when the slope was not significantly different from 1 and, simultaneously, the intercept was not significantly different from 0.

Experimental Plan: Sampling and Spectroscopic analysis

Sampling of EVOOs was performed in the context of the Violin Project (project code: 2016-0169, founded by the Ager Foundation); all the collected EVOOs were produced with olives harvested in the season 2017-2018. Sampling was planned with the aim of fully representing the whole Italian production; to this purpose, 106 samples were collected from the ten most productive Italian regions: Apulia, Tuscany, Sicily, Trentino-South Tyrol, Umbria, Veneto, Calabria, Latium, Sardinia and Liguria. The number of samples analyzed for each region is proportional to the importance of their production (in terms of quantity). This set included 28 PDO (Protected Designation of Origin) and 10 PGI (Protected Geographical Indication) EVOO samples.

In order to avoid any sample degradation, fresh olive oil samples were stored at 4 °C under dark conditions (in amber bottles) till analysis.

For determining the quality parameters of the EVOO samples, destructive analyses were performed on the whole set of EVOOs. In more detail, FAMEs were quantified using a fast-GC approach while TAGs were obtained thanks to a UHPLC system.

For FAMEs determination, samples were prepared as follows: 25 mg of EVOO sample were weighted in a 5 mL screw-top test tube. The lipid fraction was trans-esterified adding 100 µL of methanolic potassium hydroxide solution (KOH/MeOH, 2M). Thereafter, FAMEs were extracted using 1 mL of n-heptane; the reaction mixture was shaken vigorously for 30 s. After 5 min, the upper FAMEs layer became clear and

ready to be injected into the GC system. FAMEs quantification was carried out on a GC-2010 (Shimadzu, Milan, Italy) equipped with a split-splitless injector (280°C), an AOC-20i+s autosampler, and a FID detector. SLB-IL60, [1,12-di(tripropylphosphonium)dodecane bis(trifluoromethylsulfonyl)imide], 15 m × 0.10 µm df × 0.08 mm ID (Merck Life Science, Darmstadt, Germany) was operated under programmed temperature: 180°C to 230°C at 15.0°C/min. The injector was held at a temperature of 280°C; injection volume: 0.2 µL; injection mode: split 1:250. The FID temperature was set at 280°C (sampling rate 40 ms) and gas flows were 40 mL/min for hydrogen, 40 mL/min for make up (nitrogen) and 400 mL/min for air, respectively. Carrier gas was hydrogen, at a constant linear rate of 90.0 cm/s and a pressure of 606.4 kPa.

Regarding TAGs, samples were analyzed using a Nexera X2 system (Shimadzu, Kyoto, Japan), consisting of a CBM-20A controller, two LC-30AD dual-plunger parallel-flow pumps (120.0 MPa maximum pressure), a DGU-20A5R degasser, a CTO-20AC column oven, a SIL-30AC autosampler, and a SPD-M30A PDA detector (1.8 µL detector flow cell volume). The UHPLC system was coupled to an ELSD (Evaporative Light Scattering Detector) detector (Shimadzu, Kyoto, Japan). Separations were carried out on two serially coupled Titan C18 100 × 2.1 mm (L × ID), 1.9 µm dp columns (MilliporeSigma, Bellefonte, PA, USA). Mobile phases were (A) acetonitrile and (B) 2-propanol under gradient conditions: 0-105 min, 0-50% B (held for 20 min). The flow rate was set at 400 µL/min with oven temperature of 35 °C; injection volume was 5 µL. The following ELSD parameters were applied: evaporative temperature 60° C, nebulizing gas (N2) pressure 270 kPa, detector gain < 1 mV; sampling frequency: 10 Hz. NIR spectra were acquired in the transmission mode with an FT-NIR spectrophotometer (Buchi NIRFlex N-500, Flawil, Switzerland), in a module for liquid analysis equipped with six positions for sample vials. The spectral profiles were acquired in the whole NIR region, from 4000 cm^{-1} to 10,000 cm^{-1}, with a resolution of 4 cm^{-1} and 8 scans for each sample. All measurements were performed at temperature controlled directly by the instrument (35 ± 0.5 °C). Two types of cells for liquid samples were used in the present study: quartz cuvettes (QC) and disposable glass vials (DGV). QC were 100-5-40 SUPRASIL® 300 (Hellma Mullheim, Germany) rectangular-section cells, with a 5-mm optical path length, height of 45.0 mm and

volume capacity of 1.75 mL. DGV were 548-0042 transparent glass vials (VWR International BVBA/SPRL, Leuven, Belgium), with a 3-mm optical path length, height of 40.0 mm and volume capacity of 1 mL.

Samples were acquired randomly in duplicate, and the average spectrum was used for data analysis, in order to minimize unwanted spectral variability.

In more detail, EVOO samples were put into QC without any physical or chemical pre-treatment. After the analysis, to prepare the cuvette for further acquisitions, each QC was washed with a surfactant detergent in warm water, followed by acetone, and then dried. Another aliquot of the same samples was placed in the DGV and the NIR spectra were directly recorded using the same method as for QC.

Chemometric approach purposed

The whole data analysis was performed in the Matlab environment (The MathWorks, Inc., Natick, MA, USA, Version 2016b) using both the PLSToolbox package (Eigenvector Research, Inc. Manson, Washington) and in-house functions.

First, NIR transmittance spectra were converted into the absorbance scale (Log(1/T)) for a direct interpretability of the outcomes [20]. Then, a noisy region at the end of the signal and without significant absorption was removed, and the spectral range reduced from 10,000 to 4528 cm^{-1}. Subsequently, spectra were organized in two matrices containing 106 rows and 1369 columns, samples and spectral variables, respectively. The first matrix was related to the acquisitions performed with QC, while the second one contained the signals obtained with DGV.

For model development, the two data matrices obtained with QC and DGV were divided into a calibration set (including 80% of samples) and a test set (including 20% of samples) thanks to the application of the Kennard and Stone algorithm [21].

Before model computation, a comparison between eight different combinations of data pre-treatments was performed in order to select the most suitable pre-processing strategy and to improve the subsequent calibration models. The application of 4 data transformations (two column and two row pre-processing algorithms) was evaluated, considering also their combinations:

- Column mean centering
- Column autoscaling
- Standard Normal Variate (SNV) transform + column mean centering
- Orthogonal Signal Correction (OSC) + column mean centering
- SNV + OSC + column mean centering
- SNV + column autoscaling
- OSC + column autoscaling
- SNV + OSC + column autoscaling.

SNV was tested, as it allowed to correct for baseline vertical shifts and global intensity effects, typically arising from light scattering phenomena in vibrational spectroscopy [20]. OSC was evaluated in order to remove some of the information embodied in spectral data that is unrelated (orthogonal) to the quantitative variable to be modelled (Y-vector); in this way, ideally, just the useful information related to the response is maintained in the X-block [22]. Both strategies for column pre-processing (mean centering and autoscaling) were considered, alone and in combination with row transforms.

The best pre-processing combination was chosen, for each model, evaluating the root mean square error in cross-validation (RMSECV), within a cross-validation cycle with 5 deletion groups, using the venetian blind scheme.

After performing the pre-treatments optimisation, Principal Component Analysis (PCA) was applied as an exploratory tool useful to identify the presence of possible outliers in the dataset and they were not found [data not shown].

To reach the final aim of statistically comparing the prediction ability of the models built using spectra measured with QC and DGV, the Passing-Bablok regression method [23] was applied on the pairs of Y values predicted by the models developed for each quality parameter separately. This regression method is particularly suitable for method comparison, since it is a symmetrical non-parametric technique, which is able to build regression models also when the independent variable is affected by a significant uncertainty. The estimation of a linear regression equation between the two

data vectors, obtained with two different methods or devices, both measured with an error associated, allows one to statistically evaluate the similarity/diversity between the two independent estimations. To do this, slope and intercept of the fitted line were calculated, and a significance test was conducted at 95% confidence level. The null hypothesis (H0) was verified when the slope was not significantly different from 1 and, simultaneously, the intercept was not significantly different from 0.

Research outcomes

Among the variables describing EVOO quality measured with the reference methods within the Violin project (see previous paragraph 2.2), six of them, whose range of variability was less reduced than for the other quality parameters, were considered for the comparison between QC and DGV. The other FAMEs and TAGs showed a so small variability – comparable with the magnitude of the measurement error – as to hinder their use for the development of reliable multivariate calibration models. In more detail, the selected TAGs were: dioleoyllinoleoyl-glycerol (OOL), oleoyl-linoleoyl-palmitoylglycerol (OLP) and triolein (OOO), while the selected FAMEs were: palmitic (C16:0), oleic (C18:1n9) and linoleic (C18:2) acids.

A subset of 80 EVOO samples was chosen by the Kennard and Stone algorithm for constituting the calibration set, and the remaining 26 samples were used as the test set, to validate the quality of the regression models in prediction.

In order to select the most suitable strategy to pre-process the NIR spectral profiles and the optimal complexity of PLS models, for both QC and DGV data, an optimization procedure was performed. It is important to underline that independent optimizations were performed for QC and DGV data; for each variable considered (three FAMEs and three TAGs), a PLS regression model was computed. In more detail, PLS models were calculated retaining an increasing number of LVs, from 1 to 10, and applying different spectra pre-treatments, according to the list presented above. Figures 1 and 2 show the RMSECV for each of the 96 calculated models (48 on QC data and 48 on DGV data) as a function of the number of LVs; different colours are used to identify the spectral pre-processing applied. This straightforward representation allows to easily individuate the type of pre-processing and the number of LVs that, in

combination, minimizes the error of each PLS model in cross-validation (optimal complexity). Figure 1 resumes the model computation on the spectral data acquired using the traditional QC, while Figure 2 refers to the model developed for spectra coming from the DGV data.

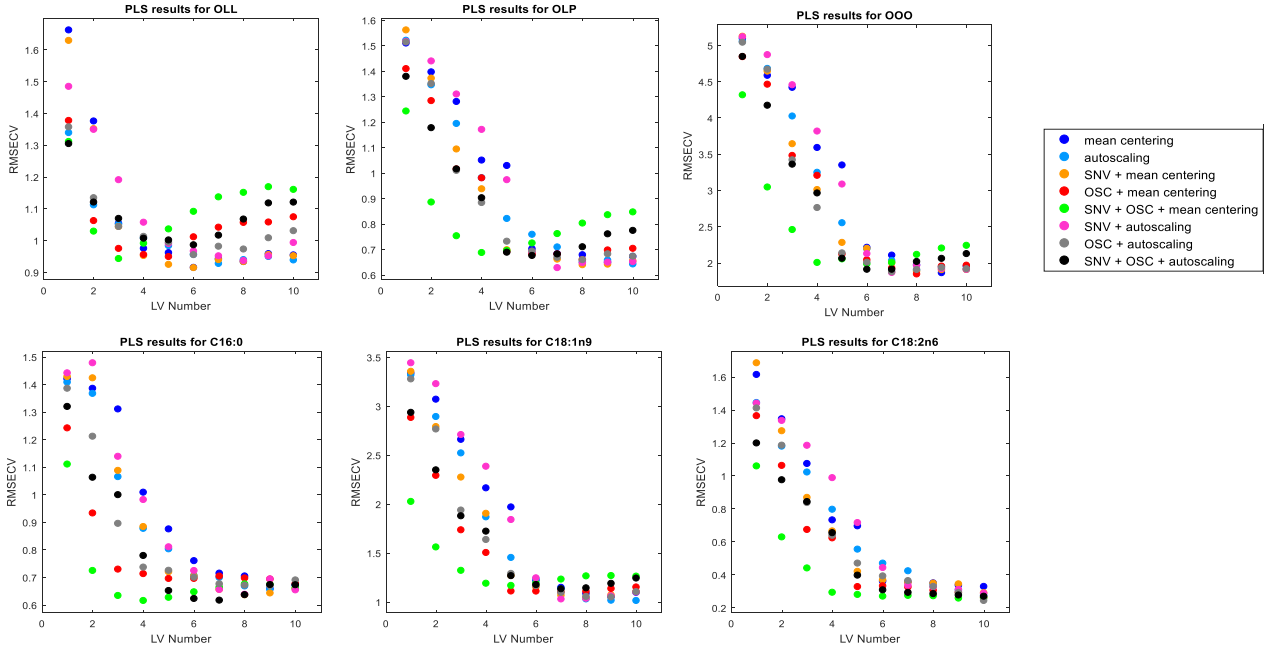

Figure 1: RMESCV of PLS regression models of NIR spectra acquired with quartz cuvettes (QC) for evaluating eight combinations of data pre-processing

For all the quality parameters modelled using the QC spectra, SNV + OSC + mean centering (represented in green in Figure 1) turned out to be the best combination, as it led to a minimum RMSECV with less complex models (lower number of LVs), if compared with other pre-processing strategies. From a global evaluation of the QC models, a LV number ranging from 4 to 6 was considered as the best compromise between model complexity and associated error (*data not shown*).

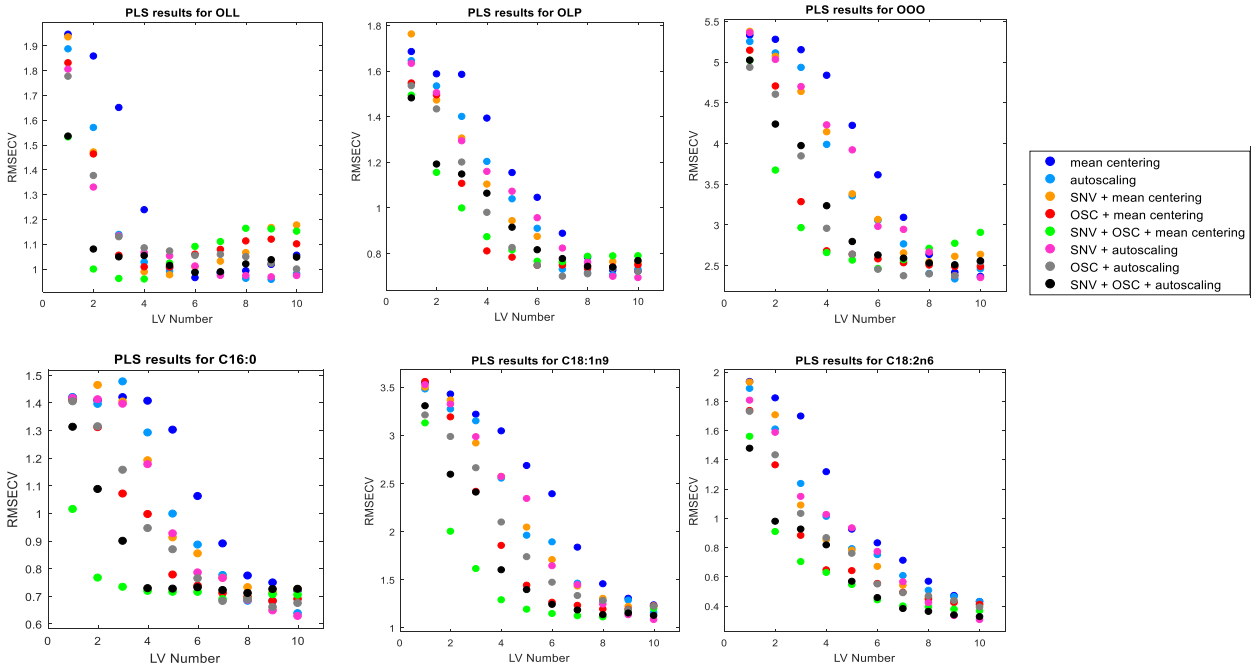

Figure 2: RMSECV of PLS regression models of NIR spectra acquired with disposable glass vials (DGV) for evaluating eight combinations of data pre-processing.

The same considerations can be drawn when considering the results obtained by the modelling of DGV spectra: for these models, the combination of SNV + OSC + mean centering (represented in green in Figure 2) has proved to be the most suitable strategy for minimizing RMSECV. To better highlight the effect of the selected combination of pre-processing on the data acquired, in Figure 3, original spectral profiles and spectra after pre-treatment, are shown: Figure 3a shows the raw signals acquired using QC, while Fig. 3b represents the QC spectral profiles transformed by SNV + OSC for variable C18:1n9. Similarly, Fig. 3c shows the original signals acquired using GDV and Fig. 3d the data transformed, for the same variable, using SNV+OSC. Using two different row pre-treatments such as SNV and OSC, it was possible not only to remove the unwanted effects caused by interferences of scattering, but also to emphasize the information embodied within the spectra according to the feature that was modelled. This approach permitted to decrease the number of LVs to retain and, therefore, the complexity of the models. For a better comparison of raw and transformed profiles, mean centering was not included in this representation.

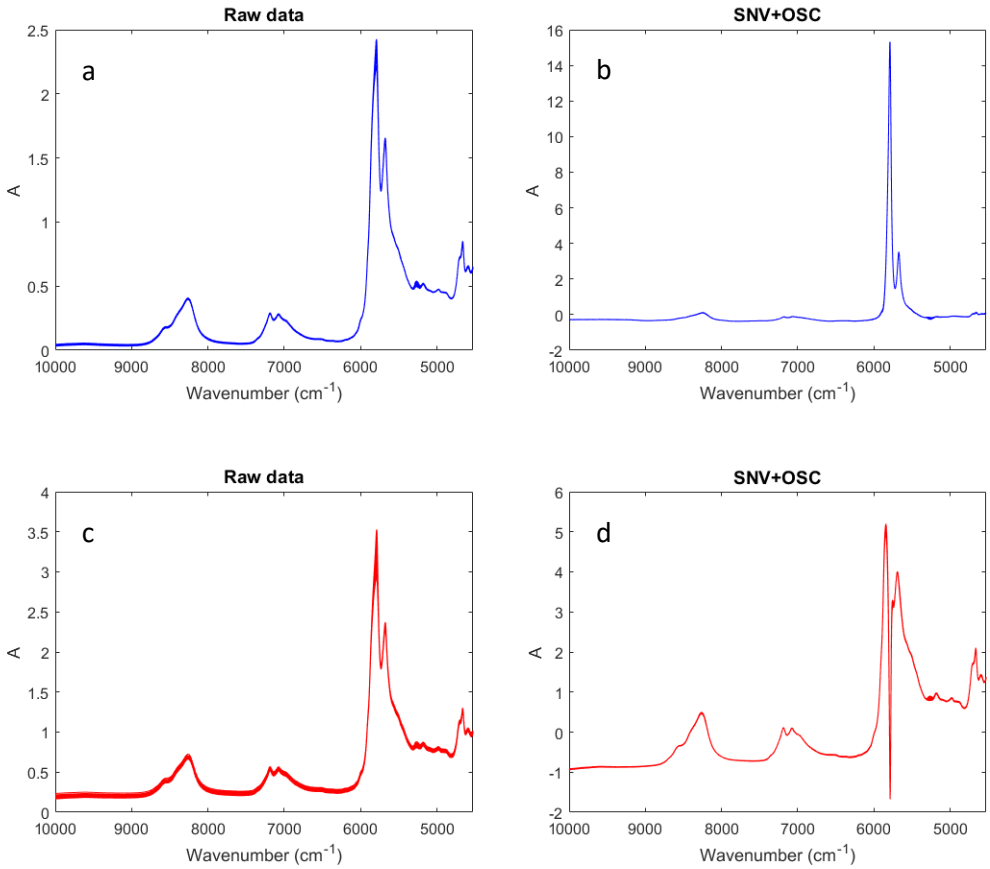

Figure 3: NIR spectra of EVOO samples before and after row pre-processing measured in QC (a-b) and DGV (c-d).

After choosing the proper data pre-treatment, PLS models were validated on samples belonging to the test set. The model parameters, calculated on pre-processed spectra, are presented in Table 1 for both QC and DGV data. For each quality parameter, a direct comparison between QC and DGV model can be performed in terms of number of LVs selected, error in cross-validation and in prediction. In more detail, the root-mean square error in cross validation (RMSECV) and in prediction (RMSEP) are reported both in the corresponding variable unit (area %) and also as percentage calculated in respect to the mean (PCRMSE). The percentage value allows a direct understanding of the model goodness.

Table 1: Calibration and prediction outcomes of PLS models for quartz cuvettes (QC) and disposable glass vials (DGV)

Quality parameter predicted (Y)	Type of cuvettes	Mean Y	Y range (min-max)	Number of LVs	RMSECV	PCRMSECV	RMSEP	PCRMSEC
OOL	QC	13.04	1.53-14.29	4	0.99	7.59%	0.75	5.75%
	DGV			4	0.96	7.36%	0.91	6.98%
OLP	QC	6.99	4.20-12.92	4	0.69	9.87%	0.68	9.73%
	DGV			6	0.76	10.87%	1.09	15.56%
OOO	QC	38.36	23.13-50.22	4	2	5.21%	1.62	4.22%
	DGV			6	2.46	6.41%	2.1	5.47%
C16:0	QC	12.86	9.53-16.40	4	0.61	4.74%	0.58	4.51%
	DGV			6	0.71	5.56%	0.77	5.97%
C18:1n9	QC	72.48	58.55-79.32	5	1.17	1.61%	1.2	1.66%
	DGV			5	1.19	1.64%	1.29	1.79%
C18:2n6	QC	7.49	4.78 -16.38	5	0.28	3.74%	0.28	3.74%
	DGV			6	0.44	5.87%	0.48	6.41%

The standard errors of the reference analyses expressed in the corresponding variable unit (area %) range between 0.40 and 1.30 and between 0.02 to 0.10 for FAMEs and TAGs, respectively; thus, for some of the models presented, the results obtained, in terms of predictive capability, cannot be considered completely satisfactory. This was verified and especially for OOL and OLP, it is mainly due to the reduced variability for these response (Y) variables in the EVOO samples of the calibration set. This did not allow to obtain PLS regression models with high predictive performances.

Looking at the results, it is possible to notice that the prediction errors are slightly – but not significantly – lower for models calculated using QC. Indeed, a numerical comparison between RMSECV% and RMSEP% of the PLS models is not the most appropriate way to understand if the predictive performances of the two types of cuvettes are effectively comparable. Therefore, to verify if the differences among the QC and DGV were statistically significant, Passing-Bablok regression was applied on the Y values predicted on test set data with both QC and DGV models. The null hypothesis (H_0) of the joint significance test was that the slope is not significantly different from 1 and that the intercept is not significantly different from 0, at a 95% confidence level; the results of these tests are presented in Table 2. For the sake of completeness, for both slope and intercept, both the limits of acceptability (LL = lower limit and UL = upper limit) and the calculated value (CAL) are reported.

Table 3: Results of joint test on slope and intercept values of the regression lines (from Passing-Bablok regression), at a 95% confidence level.

Quality parameters	Slope LB	Slope UB	Slope CAL	Intercept LB	Intercept UB	Intercept CAL	H_0
OOL	1.09	2.26	1.58	-16.39	-1.22	-7.52	H_0 accepted
OLP	1.08	1.82	1.40	-5.30	-0.41	-2.53	H_0 accepted
OOO	1.17	1.80	1.48	-30.60	-5.92	-17.83	H_0 accepted
C16:0	0.72	1.17	0.89	-2.18	3.41	1.38	H_0 accepted
C18:1n9	0.81	1.32	0.99	-23.06	14.57	1.42	H_0 accepted
dC18:2n6	1.02	2.55	1.55	-11.09	-0.28	-3.97	H_0 accepted

Outcomes of the tests indicated that there were not statistical differences between FAMEs and TAGs Y values predicted from spectra measured with QC and from spectra measured with DGV; the null hypothesis (H_0) was, in fact accepted, for all the six parameters (OOL, OLP, OOO, C16:0, C18:1n9, C18:2n6) considered, at the 95% confidence level. Considering these results, it was possible to state that comparable results were obtained for the prediction of the EVOO quality parameters considered, with both quartz cuvettes and disposable glass vials.

Conclusion and scientific impacts

In this study, with the final aim of reducing the time of NIRS acquisition for olive oil 'highly frequent' analysis, a critical comparison between analytical performances of QC and DGV, based on the determination of parameters which affect olive oils quality (FAMEs and TAGs), was performed.

In more details, a large set of EVOO samples was analysed by NIRS using both QC and DGV, and the spectra were used to build PLS calibration models for predicting some EVOO quality parameters.

Optimisation of pre-processing showed that a joint application of SNV + OSC + mean centering led to the best models in terms of prediction error in cross-validation and model complexity (lowest number of LVs).

Thanks to a significance test based on Passing-Bablok regression, it was possible to highlight that there are not statistical differences between models calculated with QC and those obtained with DGV; this statement was demonstrated for all six the parameters (OOL, OLP, OOO, C16:0, C18:1n9, C18:2n6) considered.

The present study demonstrated that the employment of DGV instead of quartz vials reduces time and costs for the acquisition of NIR spectra, representing a key point for the automation in the olive mill industry. The challenge of implementation of NIR equipment in the production chain is, in fact, of great concern to the olive oil industry, and it depends on the availability of dedicated analytical solutions able to provide an accurate multicomponent analysis in a short time and with little effort and, in this context, the use of DGV allows to eliminate the washing and drying times necessary for the QC.

In order to understand if DGV can replace QC also for different analyses, this study should be extended, applying the same scheme to data coming from the prediction of other parameters of interest.

2.2 ANALYSING THE WATER SPECTRAL PATTERN BY NEAR-INFRARED SPECTROSCOPY AND CHEMOMETRICS AS A DYNAMIC MULTIDIMENSIONAL BIOMARKER IN PRESERVATION: RICE GERM STORAGE MONITORING

Scientific Background and aim of the work

Water activity is one of the fundamental concepts on which food processing and food storage are based, extensively investigated since the beginning of the eighties and accepted in many national and international food legislations. Water activity is defined as the partial vapor pressure of water in the food matrix under study in respect to the standard state partial vapor pressure of pure water; it is a measurement of the energy status of water in a system. Thanks to this definition, it is possible to understand the importance of such parameter - it represents the amount of water in a food product available for biochemical reactions and it is an indicator of food stability with respect to microbial growth. Moisture content is another important parameter for stability of food during storage – it is a measure of the quantity of water in the product, but water activity provides information about which part of this water is available for chemical reactions. Nowadays, the non-linear relationship between water activity and moisture is accepted worldwide and schematized in the moisture sorption isotherm curves [24]. These isotherms are substance and temperature specific, and they are useful in predicting product stability over time in different storage conditions. Although, water activity in low-moisture foods depends on the storage conditions such as relative humidity and temperature, it is generally accepted that at values of water activity below 0.60 threshold microbial spoilage is not likely to occur [25], [26]. Control of the water activity and moisture content is the basic principle of food preservation and prevention of microbial and chemical deterioration – both are aimed at actually manipulating the water, either by removing or binding [26]. Water content and water molecular structure are therefore the crucial factors determining the stability and modifications of products during storage. Despite being microbially stable, the food during storage may undergo

changes due to the chemical and enzymatic reactions [26], which are dependent on the water activity and can influence the properties of interest for consumers such as taste or texture. Several methods have been proposed for measuring water activity, from resistive electrolytic or capacitance hygrometers to dew cells and others. With both methods, vapor–liquid equilibrium must occur in a sample chamber in which the temperature is continuously measured. There are also attempts of developing novel type water activity meters based on electromagnetic spectroscopy which could provide faster, no contact and non-destructive measurements [27], [28].

The role of water activity in determining the reaction rate is demonstrated to be crucial, in particular, for low-moisture food. Labuza in 1970 [29] proposed a global food stability map in which the evolution of crucial reactions that occur in food products is described, in respect to the water activity level. In particular, in food matrices with a moisture content lower than 25%, lipid oxidation is the major cause of quality degradation in food [30]. This is the case for rice germ, a by-product of the milling industry, nowadays considered as a novel food because of its interesting nutritional value. Rice germ is the embryo part of rice and represents 1–3% of rice kernel (depending on the variety), but it contains all the nutritional elements needed for the growing of a new plant [31]. In commercial products, it is completely removed during the milling process in the whitening phase. The reason is the high content of unsaturated fatty acids in rice germ, which are very susceptible to oxidative rancidity, strongly limiting the storage stability and shelf life [32].

In light of these considerations, in the present study, the storage of rice germ at different water activity levels was chosen as a case study for understanding the water molecular structure within the rice germ and how it is related to the water activity and the changes during storage. Indeed, despite its importance and its practical implications in food storage, in-depth studies of how the water activity is related to the water molecular structure are scarce and many phenomena are still left unexplained. For example, it is still unclear why at the same water activity level some foods are stable, but others are not. To understand the mechanism of food preservation/degradation, the knowledge of the water structure is necessary, because its mobility and availability for biochemical reactions depend on the type of interactions with other food components [33].

In order to fulfil the objective of this study, the aquaphotomics approach was chosen – an approach to study water by focusing on its interaction with electromagnetic radiation. Aquaphotomics is an "omics" discipline, established by Roumiana Tsenkova [34] and thoroughly presented in recent review [35], [36]. The main subject of this new approach is to understand the integrative role of water in biological and aqueous systems by monitoring how the water spectrum changes under various perturbations. The near infrared (NIR) wavelength region of the electromagnetic spectrum, in particular, the first overtone of the OH stretching vibrations (1300 – 1600 nm) was shown to be an excellent window for the observation of the water molecular structure and, in addition, enables non-destructive measurements. In contrast to the traditional NIR spectroscopy studies, in which the water absorption band is considered as masking the real information, aquaphotomics considers the water spectral pattern as the main source of information, like a sensor or a mirror of what is happening in the samples under study. This principle of indirect measurements is described as a "water mirror approach [34], [35]. In more detail, aquaphotomics studies have been done in specific spectral regions when a certain perturbation of interest is applied on samples under study; in these regions, specific water absorptions can be found with the highest probability, defining the water matrix coordinates (WAMACS) [34], each one imputable to a specific water molecular formation (water molecular species). The bands showing highest spectral variations considered as "activated", within the well-established WAMACS, are monitored for a comprehensive understanding of the system. The combination of these activated bands defines the water absorbance spectral pattern (WASP), which describes the condition of the whole aqueous system related to its functionality (i.e., how the system behaves under the specific perturbation or its properties). The representation of the WASP is done by using special radar charts, called aquagrams [37]. This rigorous multivariate approach allows to standardise the data processing procedure, fundamental for performing any type of comparison or generalisation of the aquaphotomics results.

Thanks to this novel approach, the dynamics of various water molecular structures related to different initial water activities in rice germ along the storage were investigated, laying the first stone for understanding the water molecular structures responsible for the phenomenon of water activity in general. The present work adds a

fundamental knowledge to the understanding of the complex nature of the processes that biological matrices undergo during storage and the role of water molecular structure in it.

Experimental Plan: Sampling and Spectroscopic analysis

The experimental plan was designed in collaboration with Società Agricola Cooperativa Rondolino (Livorno Ferraris, Vercelli, Italy) which also provided rice germ samples, separated from the bran through sifting to a purity degree of about 85%. Rice germ samples were stored at 27 °C, at three different levels of water activity (a_w) and three different storage atmospheres (SAP) (with the exception of a_w = 0.55 sample, which was only stored in air atmosphere), for a total of 7 combinations, as shown in Malegori et al. 2020 [45] and summarized in Table 1.

Table 1: Experimental material and conditions. Sample identification, reporting: sample code, water activity, moisture degree and storage atmosphere (SAP) (at 27 °C).

Sample ID	a_w	Moisture content (g/100 g)	SAP
A1	0.55	9.71 ± 0.11	Air
B1	0.45	7.82 ± 0.15	Air
B2	0.45	7.82 ± 0.15	Vacuum
B3	0.45	7.82 ± 0.15	Argon
C1	0.36	6.86 ± 0.08	Air
C2	0.36	6.86 ± 0.08	Vacuum
C3	0.36	6.86 ± 0.08	Argon

Rice germ samples were packaged in cans (210 mL, filled with 130 g of sample) and then stored for 320 days. Samples at three levels of a_w were first analyzed before packaging (time t_0). Then, during the 320-day storage period, packaged rice germs were sampled and analyzed seven times. The a_w levels were monitored for each sample along the whole storage period and did not show any significant modifications during the entire observation period, as it was detailed in a previous study [45].

To investigate chemical modifications of rice germ during storage according to the level of a_w, a robust rice germ NIR spectral data analysis procedure, based on an aquaphotomics approach [36], was developed and the workflow is schematically presented in Fig 1

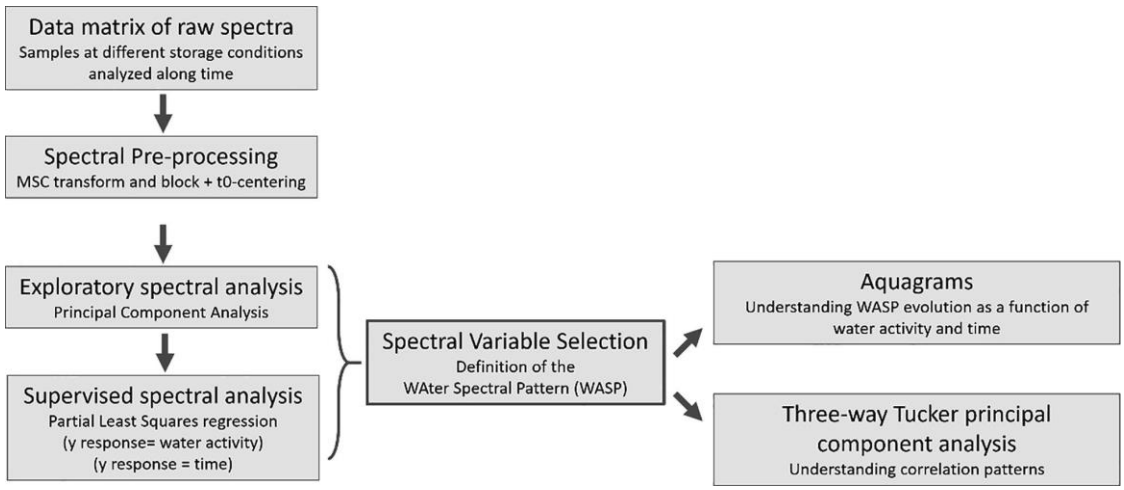

Figure 1: Spectral data analysis workflow. To investigate chemical modifications of rice germ during storage according to the level of aw, a robust rice germ NIR spectral data analysis procedure, based on an aquaphotomics approach, was developed and executed according to the present workflow

Aquaphotomics analysis was performed using only the first overtone of the O-H stretching vibrations, focusing on the water absorption pattern in the NIR region from 1300 to 1600 nm.

Rice germ samples were analyzed by near-infrared (NIR) spectroscopy in the reflection mode (data recorded as Log (1/R) or pseudo-absorbance), using a Fourier Transform (FT) NIR spectrometer (MPA – Bruker Optics, Milan, Italy) equipped with a rotating sample holder; spectra were acquired in the range 12,000 – 4000 cm^{-1} (800

– 2780 nm), with a 4 cm^{-1} resolution, and 64 scans for both samples and background. To obtain a representative spectrum of each sample, the whole content of rice germ coming from each can was placed in a glass Petri dish at room temperature (20 – 22 °C) before performing the rotational acquisition. To perform the data processing according to the aquaphotomics approach, it is fundamental to standardise accurately the acquisition step, in order to minimise the experimental noise and the influence of unwanted perturbations on the spectral profile.

Chemometric approach purposed

The first step in data evaluation was the pre-processing, aimed at smoothing differences between samples stored at the three different atmospheres, in order to focus on spectral features, characteristic for variations in a_w. As the row data pre-processing, a multiplicative scatter correction (MSC) was applied, according to the aquaphotomics procedure, allowing us to properly address the absorption bands without taking into account global intensity effects [20]. A block-wise centering was performed by subtracting the average profile of spectra recorded on samples stored at each modified atmosphere (considered as a block) from spectra of samples at the same condition. This set of pre-processed spectra will be from now on referred to as BW-centered (block-wise) spectra. To account for another important factor of the data set under study – the time trend – another specific data pre-treatment was developed, trying to decompose the variability related to a_w and the one related to storage time. To do this, spectra were also centered with respect to the time 0, by subtracting the spectrum obtained at t_0 (before packaging) from all of the spectra subsequently recorded for the samples at the corresponding a_w. This pre-treatment procedure minimised original differences among batches, and from now on will be referred to as t_0-centering.

Data pre-treated by these two strategies were submitted to the subsequent exploratory processing step (Principal Component Analysis – PCA), with the aim of understanding the contribution of the two factors, a_w and storage time on rice germ modifications. PCA is a well-recognised and informative method that explores data structures without

using a priori information [92]. This exploratory step was performed on both data sets, in order to understand which variables are more accountable for sample groupings, according to a_w level and time trend. This approach was possible thanks to the joint interpretation of the score and loading plots, graphical outcomes of PCA analysis.

To emphasise the information related to a_w and to test the consistency of the PCA outcomes, a supervised method was applied on pre-processed data. Partial Least Squares (PLS) regression analysis [19] was performed, using a_w level as the response variable (y variable). Evaluation of PLS coefficients allows insight into which variables of the predictors (x matrix) have the highest contribution for successful differentiation between samples in terms of the water activity.

To confirm the consistency between the two approaches, highlighted variables are expected to be the same as those identified by the exploratory and the supervised method. Both steps aimed at highlighting the most informative NIR wavelengths representing respective water molecular structures, the ones that are activated by the variations in a_w and storage time. Getting to the heart of the aquaphotomics approach, the wavelengths selected with the previous steps, are used to define the Water Spectral Pattern (WASP) [34] for the study of the rice germ evolution during storage. The variations of the absorption at these selected wavelengths were graphically presented using aquagrams [36], [38] allowing an integrative interpretation of the phenomena under study. Two types of aquagrams are presented in the current study: the first one is aimed at enhancing the differences between rice germ samples stored at different a_w levels; the second one is focused on variations of samples along time, split according to the levels of a_w. In both cases, information related to different SAPs was minimised by application of BW-centering. It is important to underline that the pseudo-absorbance (Log(1/R)) values were range scaled (between the minimum and the maximum values of the subplots within each figure) for a direct comparison of all the aquagrams presented on each figure. What becomes evident from the procedure described is that there is a strong interaction between a_w levels and a time trend, which is very difficult to decompose. For this reason, the last step of the data processing is focused on understanding such a pattern of correlation, thanks to a three-way data analysis [95]. Data were submitted to Tucker3 decomposition after j-scaling (autoscaling performed along variables) [96]. Performing Tucker3 PCA on the t_0-

centered data, in fact, allows to represent, in a unique orthogonal space (triplot): a_w levels, sampling times and the selected wavelengths.

Exploratory analysis: spectral averaging and principal component analysis (PCA)

The results of the first steps of the analysis, presented in Fig. 2a, show the average spectra of the samples at different water activity levels after the pre-processing steps (multiplicative scatter correction, BW and t0-centering). The role of pre-processing was two-fold: by BW-centering the differences among different storage atmospheres, (SAP, codified by numbers: 1, 2 and 3) were minimized, while the t0-centering minimises, in addition, the differences among a_w batches (A, B and C) before packaging, at time t_0. As a result of this, a clear difference between the spectral profiles can be observed for spectra of rice germs stored at different a_w levels.

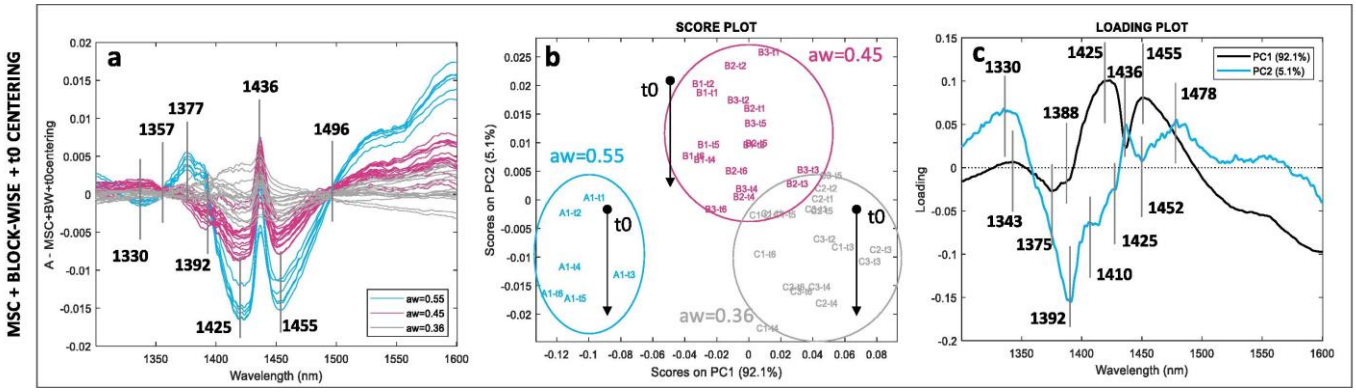

Figure 2: Results of exploratory analysis. (a) Spectral profiles after preprocessing using MSC, block-centering and t0-centering; (b) PCA score plot of the preprocessed spectra, (c) loading plot of the preprocessed spectra. All the figures presented are codified by the same color, according to the water activity level: light blue for the highest level of a_w = 0.55; pink for the intermediate level of a_w = 0.45, and grey for the lowest level of a_w = 0.36.

Fig. 2b and 2c show the score and loading plots of PCA for the pre-processed data, respectively; the two lowest-order principal components (PC1 and PC2) account for a total of about 97.2% of the total information. The scores plot (Fig. 2b) shows that PC1 (explaining the 92.1% of variance) is associated with a_w levels and the separation among the three levels (A, B and C) is well-defined. A, B and C samples are distributed

along PC1 with the highest level of a_w (A) located at the negative score values, and the lower levels located at higher scores. Looking at the loading plot of the PC1 (Fig. 2c) it shows that different a_w values of rice germ influence the absorbances at 1343 nm, 1375 nm, 1388 nm, 1425 nm and 1455 nm, the last two bands having higher absorbance for lower aw values, while the first three listed bands characterize high a_w sample. The PC2 component (accounting for the remaining 5.1% of the variance) is separating the A and C scores from B scores, and in addition, it is possible to observe that PC2 scores are associated with the time trend. For each A, B and C group of scores, a time trend is detectable along PC2: the first sampling points (t_0 and t_1) are located at the highest PC2 scores, while the last ones (t_5 and t_6) at the lowest PC2 scores for each of a_w level score groups (trend is represented by arrows in Fig. 2b). The loading of the PC2 shows the following important wavelengths: at 1330 nm, 1392 nm, 1410 nm, 1425 nm, 1436 nm, 1452 nm and 1478 nm. While it is not clear at this point why the B scores are separated from A and C scores along the PC2, the time trend is clearly associated with the changes in absorbance at the mentioned wavelengths, most probably it indicates that time-dependent changes in water molecular structure of rice germ are not in linear relationship with the initial a_w of the samples, which obviously dictates the position of the scores at the beginning of storage period. This further means that PC2 also explains a small part of the total variance due to a_w, which translates into non-linear changes in absorbance at the bands found in loading of the PC2. Similarly, a fluctuation of scores corresponding to the sampling period within the last three months can be observed, indicating also a not completely linear relationship with time.

Supervised analysis – partial least squares (PLS) regression

To further the investigation of which absorbance bands are related to the phenomena under study and how, a supervised methodology based on PLS regression was applied. Using pre-processed spectra as the predictors, the regression modelling was performed firstly, using a_w as the response variable, and secondly, using the time as a response variable. In the case of the regression using time, the models were built

independently for each of the different a_w levels sample groups, since the previously performed PCA analysis clearly revealed differences depending on the initial a_w levels.

These models were not built for predictive purposes, but for a supervised exploration and a better understanding of the most informative spectral bands concerning a_w and the time trend. The results of PLS regression are presented in Figures 3 and 4. Fig. 3 presents PLS analysis results obtained using a_w as the response variable while Fig. 4a, b and c show agreement between values measured and predicted by PLS regression for time, at $a_w = 0.55$, $a_w = 0.45$ and $a_w = 0.36$ levels, respectively. Lastly Fig. 4d shows the regression vector coefficients for the three time regression models, together.

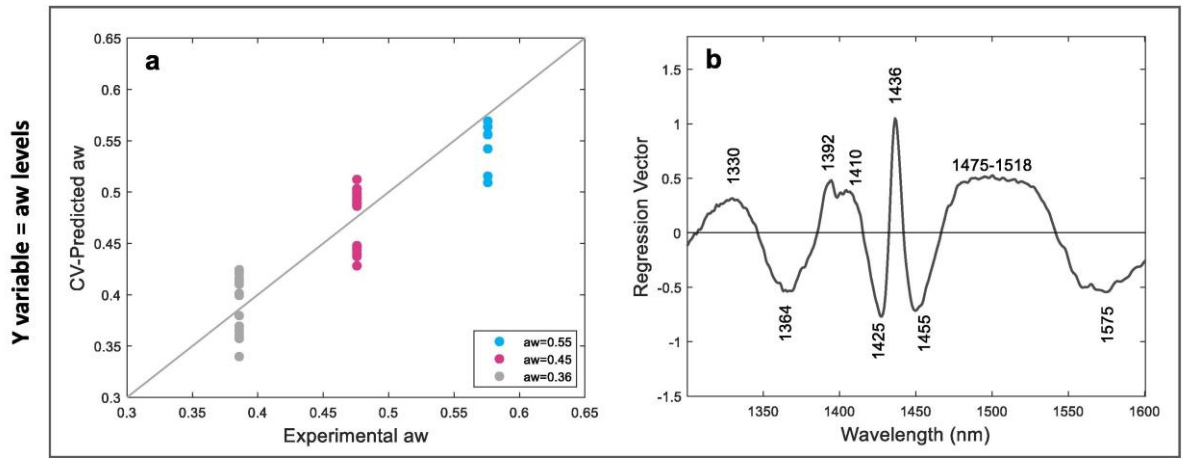

Figure 3: PLS regression analysis using a_w. PLS regression analysis on pre-processed spectra using a_w levels as dependent variable: (a) predicted vs. *experimental a_w, (b) PLS coefficients using different levels of a_w as dependent variable.*

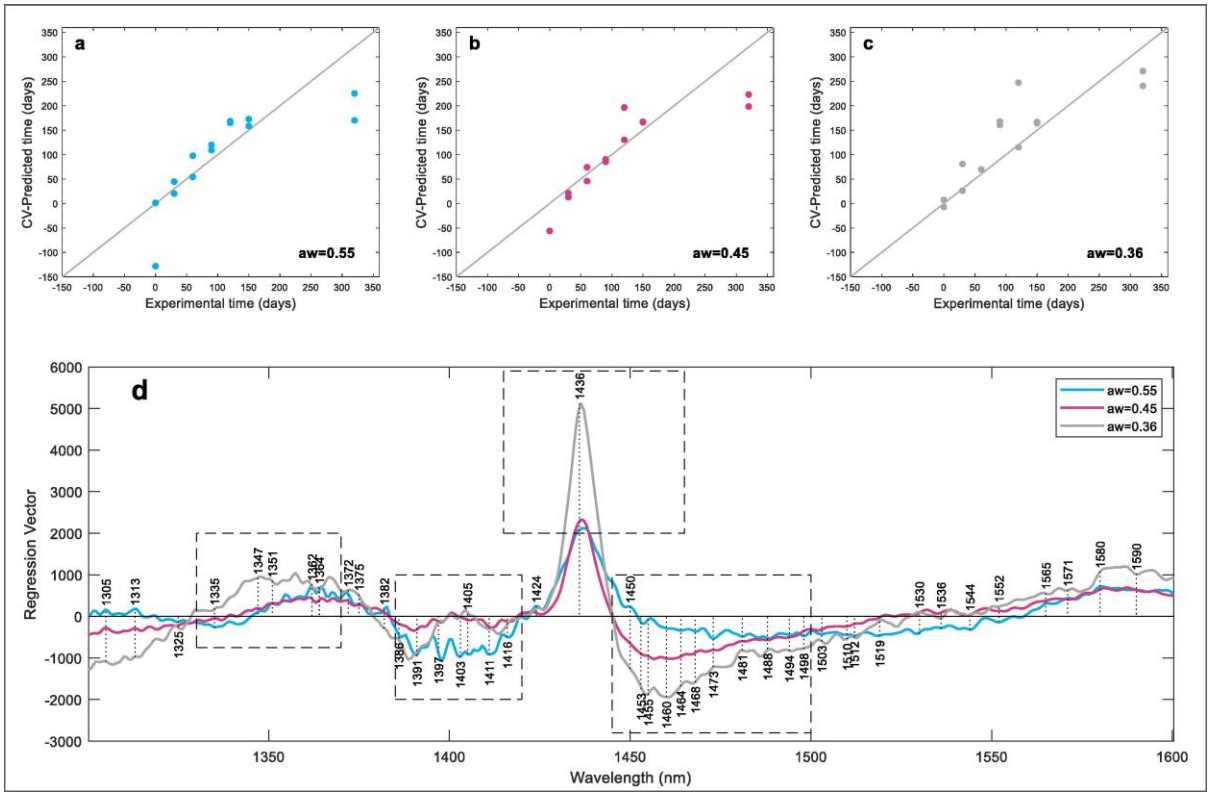

Figure 4: PLS regression using time. PLS regression analysis on pre-processed spectra using time as the dependent variable: (a) predicted vs. *experimental time at a_w = 0.55, (b) predicted* vs. *experimental time at a_w = 0.45, (c) predicted* vs. *experimental time at a_w = 0.36, (d) PLS coefficients at three a_w levels.*

Fig. 3a shows the predicted vs. experimental values in cross-validation: due to the fact that a_w levels are discrete, the vertical dispersion of scores along the same a_w value is related to model error. Although a noticeable dispersion is evident at each level, it is important to underline that no overlap is detectable between the different levels of a_w. The most important outcome of this step is reported in Fig. 3b that represents PLS regression coefficients of each spectral variable (wavelength). These values indicate the contribution of absorbance bands at each wavelength in the modelling of a_w response: the larger the absolute value, the higher the contribution of the absorbance at this wavelength (positive or negative). Some of the wavelengths highlighted (1330 nm, 1392 nm, 1410 nm, 1425 nm, 1436 nm, 1455 nm and ~ 1478 nm) are consistent with the ones observed in loadings of the PCA analysis, strongly confirming their

importance. PLS regression also revealed other important bands at: 1364 nm (negative contribution) and contribution of a broad band around 1475–1518 nm (positive contribution) which is a result of several absorbance bands whose small peaks can be observed. From the Fig. 3a it can also be observed that the modelling of a_w values slightly deviate from linearity, a finding also seen from the PCA analysis (Fig. 2b, the trend along PC2 axis). In the regression vectors of PLSR models developed for the time, for each level of water activity separately, it can be seen that the differences in regression coefficients occur in the following spectral regions 1340 – 1360 nm, 1390 – 1415 nm, 1445 – 1500 nm and in particular at one absorbance band − 1436 nm. This finding indicates that changes in the water structure during storage are different depending on the initial a_w of the rice germ when it was stored, and it dictates which water molecular species are going to reorganize. In other words, the initial a_w level upon storage defines the initial water molecular structure, i.e. the representation of particular water molecular species, which further governs possible structural changes during storage. In all the regression vectors from both time models and a_w model, common bands can be observed: 1364 nm, 1390–92 nm, 1405–14 nm, 1436 nm, 1453–57 nm and 1575–1580 nm.

Aquaphotomics analysis

Based on the analysis so far, it became evident that the bands found important for the description of the influence of initial a_w, as well as the progression of time on the rice germ samples, belong to previously well-defined water matrix coordinates (WAMACS) – absorbance bands of water found within the spectral region of the first overtone of the O-H stretching vibrations, corresponding to absorbance of different water molecular species [34]. This region has been thoroughly investigated leading to the experimental findings, theoretical confirmation and finally, definition and systematization of the mentioned 12 WAMACS into discrete intervals, each one with a specific assignment in the terms of the water molecular structure [34]. In this study, the wavelengths, revealed as important by the multivariate analysis described above, represent activated water absorbance bands [36] and majority fits within the 12 WAMACS intervals. The phrase “activated water absorbance bands” means that water

molecular species which absorb light of this wavelength are particularly affected by differences in a_w and/or duration of storage (time). The activated water absorbance bands, according to the aquaphotomics analysis protocol, can be used to visually represent the water spectral pattern (WASP) of the studied samples in special charts – aquagrams [36], [38] which allow easy insight into the molecular composition of water within the rice germ and how it changes in respect to the studied perturbation (a_w, time). In addition to the bands repeatedly found in the analysis, in order to take into account also water structures not revealed by the previous steps, the complementary wavelengths from other WAMACS intervals were also included in the aquagram representation (1343 nm, 1382 nm, 1474 nm and 1518 nm) resulting in the definition of the following WASP for studying the evolution of rice germ samples during storage: 1343 nm, 1364 nm, 1375 nm, 1382 nm, 1392 nm, 1410 nm, 1425 nm, 1436 nm, 1455 nm, 1474 nm, 1492 nm, and 1518 nm. Absorbance values recorded at these 12 wavelengths were used, after a global normalization, to build two types of aquagrams: Fig. 5 shows the profiles of each rice germ sample, colored according to the corresponding a_w level, for different sampling times, while Fig. 6 shows the profiles of rice germ samples, colored according to the different sampling times, for different a_w levels.

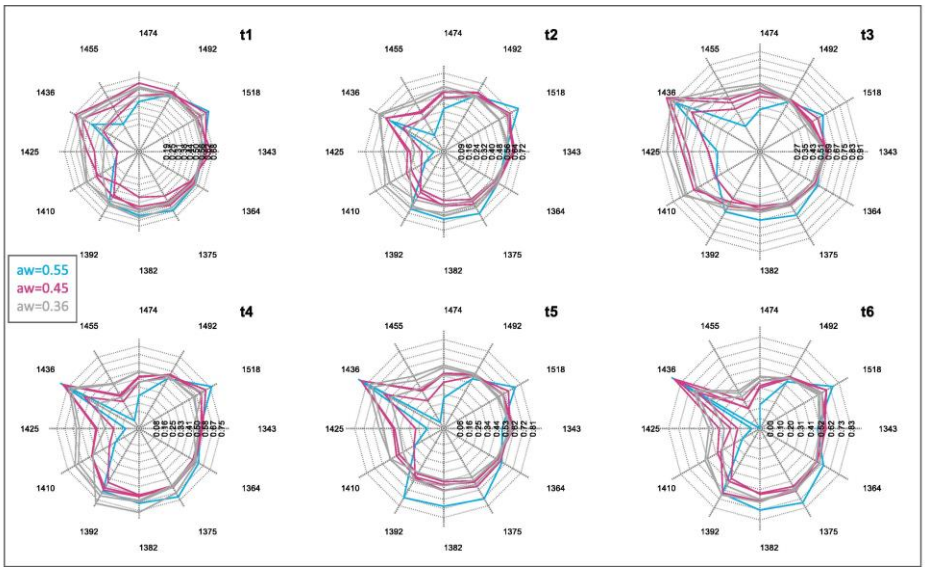

Figure 5: Aquagrams grouped according to the sampling times. The aquagrams show the differences of the samples stored at different initial aw levels at each sampling time (sampling times from t1 to t6). The profiles of rice germ samples are colored according to the corresponding aw level.

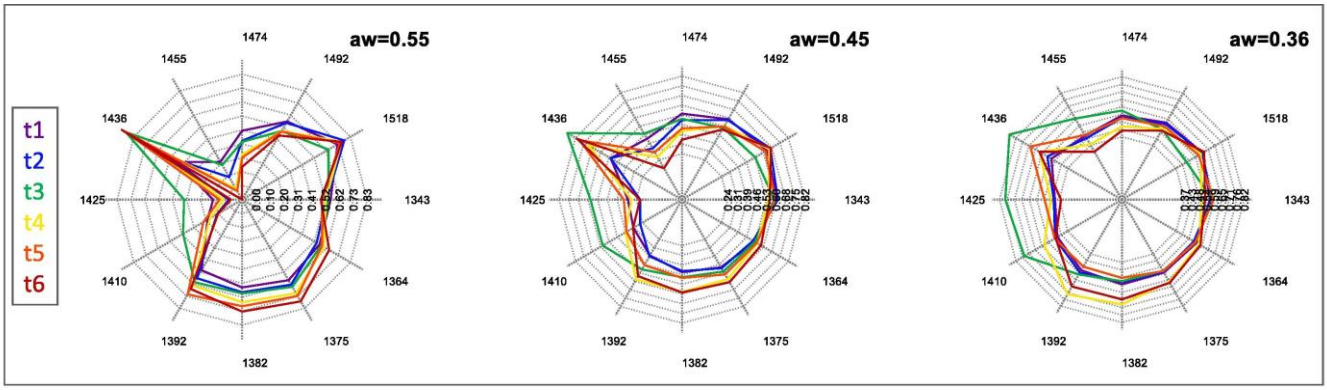

Figure 6: Aquagrams grouped according to the a_w level. The aquagrams show how each sample group, depending on the initial water activity, is changing over time (the profiles of averaged rice germ sample are colored according to the sampling time). Each aquagram is an integrative biomarker for rice germ samples at the specific water activity level.

First, let us consider the aquagrams in Fig. 5 which represent water spectral patterns (WASP) of the rice germ stored initially at different aw levels and how they compare at each measurement time point. The sample stored at the highest a_w = 0.55 level, compared to the others, throughout the storage period, shows highest absorbance at 1518 nm, and lowest at 1410 nm, 1425 nm and 1455 nm.

For this sample, the absorbance at 1436 nm increases steadily over the storage period. Also, the absorbance at three wavelengths 1364 nm, 1375 nm and 1382 nm increases, and it is especially higher when compared to the other samples. On the other hand, the samples stored at the lower levels of a_w = 0.36 and a_w = 0.45 show higher absorbance at wavelengths from 1392 to 1492 nm (with the exception of 1436 nm). Then, during the storage, as the time progresses, the absorbance in this region decreases, leaving only high intensity at 1436 nm. At the end of the storage period, samples stored at the lowest initial aw levels show the lowest value of absorbance at this band. However, despite the decrease in absorbance, the values at 1410 nm and 1425 nm remain high and higher in comparison to the sample stored at the highest a_w level. One more interesting feature for WASPs of all samples is that, despite the changes as a function of time, the absorbance at 1492 nm is where all the samples are the least different. It is the only absorbance band where WASPs of all samples, regardless of a_w or storage atmosphere, almost coincide at each measurement point along time. Among the observed absorbance bands, the 1410 nm and 1518 nm are the most well-known absorbance bands of water in the scientific literature and can be

assigned to free water molecules and strongly bound water, respectively (11). Analyzing the aquagrams in Fig. 6, which show differences in WASPs for each a_w level separately, over time, a certain trend can be detected in agreement with what was observed from the earlier aquagrams: samples at all a_w levels present a strong absorption at 1518 nm, and a weak one at 1410 nm and 1425 nm. This interesting feature is preserved along time, indicating that the ratio of these water structures is important during storage. On the other hand, for each a_w level, the major variations during storage time, and the highest absorbance can be noticed at 1436 nm, indicating the importance of the water species absorbing at this wavelength and that they vary during storage. It can also be seen from these aquagrams that the variations happen along time at the water bands 1392 nm, 1364 nm, 1375 nm and 1382 nm, and that they follow the time trend in the case of the samples with the highest a_w. The first out of these four absorbance bands can be assigned to the water confined in the local field of ions, also called trapped water or dehydration band [34], [35], [39], [40], while the latter three correspond to proton hydrates [41], [42], [43] and/or water solvation [11], while together they can be attributed to water vapour [21]. From these aquagrams also, it can be observed that absorbance at 1492 nm only decreases with time, and that during the third month of storage (t_3) all the samples show substantial change in spectral pattern at 1410 nm, 1425 nm and 1436 nm. Previous study using NIR spectroscopy and electronic nose, also indicated differences between the first three (considered still fresh) and the last three months of storage in rice germ (long time of storage).

In summary, from Fig. 6. one can notice that a certain, specific water spectral pattern (WASP) is associated with the particular a_w level of rice germ. This means that an aquagram as a graphical representation of WASP can serve as an integrated biomarker of rice germ state and as it will become clear later on, it provides much more information about not just water activity but other properties of the samples related to water structure as well.

Comprehensive evaluation - Tucker3 Principal Component Analysis (PCA)

From the previous analysis, the interdependence of a_w levels and the time trend became evident, but although certain absorbance bands that can be attributed to particular water species were recognized, how they relate to each other is still not easily decomposable. For this reason, a 3-way exploratory analysis was performed by the Tucker3 algorithm, as a tool for a comprehensive evaluation of these interactions. Tucker3 PCA was performed on the data matrix organized as a 3D data cube, in which rice germ samples, selected water absorbance bands and storage time were considered as objects, variables and conditions, respectively. The results are reported in a triplot in Fig. 7. The two lowest-order components are shown using two Cartesian axes, explaining 62% of total variance (information). Axis 1 shows that it is mainly associated with a_w variations showing a progression from lower values (left sector) to higher values (right sector). Additionally, it can be noticed that all the samples which are stored in the air atmosphere, independently of the initial aw are located on the positive side of the Axis 1, while samples stored in modified atmosphere (vacuum and argon) are on the negative. It also becomes clear from the graph that Axis 2 explains time evolution: shorter storage times are located at the negative values of Axis 2, while longer storage times (t_4, t_5 and t_6) at positive values. Regarding the NIR variables (wavelengths corresponding to activated water absorbance bands), this approach confirms once more the importance of already found wavelengths in the previous steps of the analysis.

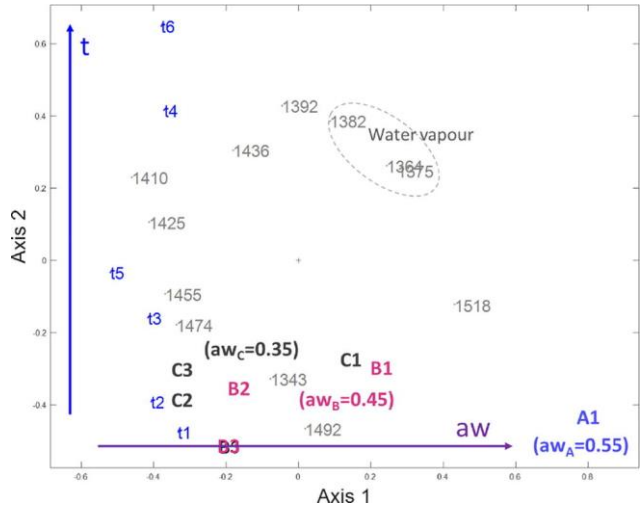

Figure 7: Tucker3 PCA triplot on pre-processed spectra. Decomposing patterns of correlation: a_w levels, sampling times and the selected wavelengths.

In particular, it is possible to clearly detect the importance of absorbance at 1410 nm and 1518 nm in defining the two extreme a_w levels; as a confirmation of the aquagrams outcomes is it possible to notice that 1410 nm band is associated with samples at low a_w level while 1518 nm with the highest level. The two other bands at 1425 nm and 1455 nm, in agreement with the PCA results, show negative correlation with the a_w, with high importance for the samples stored at lower a_w levels. Regarding the time trend, the key role of absorptions at 1392 nm and 1436 nm is evident thanks to their position in the triplot at high values along Axis 2: the "older" the rice germ (t_4, t_5 and t_6) the higher the absorption at 1392 nm and 1436 nm. The graph also shows the existence of difference between the first three and last three months of storage, where in the last three months there is a fluctuation along time, whereas t_1, t_2 and t_3 show regular increase. On the opposite end of Axis 2 is 1492 nm absorbance band, showing higher value for samples at the beginning of the storage life (t_1, t_2) while its contribution decreases along time. Moreover, 1492 nm is not related to a_w because it is located very near zero point at Axis 1, in agreement with the observations based on the aquagrams. The next important wavelengths, very influential for both storage time and aw are 1364 nm, 1375 nm and 1382 nm, located in the right-upper corner of the triplot.

As already mentioned, absorbance at these wavelengths can be attributed to solvation and protonated water, which either hydrates protons or other ions, while together they were found to represent spectral pattern of water vapour [44]. From the Tucker3 plot, it can be seen that absorbance at these bands is positively related to the time trend and a_w, and also characterizes the samples stored in the air atmosphere. The presence of water vapour, in addition to strongly bound water (1518 nm) and trapped water (1392 nm) is important for samples stored at higher initial a_w and in air. Thanks to the 3-way data processing, it was possible to confirm the outcomes of the previous data analysis in a very consistent way, highlighting correlation patterns between variables and decomposing the contribution of each mode in explaining the system under study.

Research outcomes

The presented results showed clearly that during storage, and depending on the initial a_w level and storage atmosphere rice germ undergoes changes, reflected in the changes in intensity of NIR absorbance spectra at several absorbance bands, all of which can be assigned to vibrations of particular water molecular species [34], [35], [39], [41], [42], [43], [44], [46], [47], [48] (Table 2). Some of the absorbance bands found important, correspond to water molecular species whose functionality and the role they have in food preservation is quite well-known and understood. Table 2 shows that the absorbance bands found in this work, were also found to be important for the prediction of parameters like moisture content, water activity, hardness and others. They also featured in the water spectral patterns found to be related to the certain attributes of the examined systems like mechanical and textural properties, as well as functionality (infection, stress response, viability, etc.)

Table 2: Assignments of water absorbance bands. Description of the main WASP and their importance in the understanding of the influence of initial water activity on modifications of rice germ during storage. Concordance (positive or negative) with aw and time trend. Assignments of the water absorbance bands are based on several sources.

Wavelength (nm)	Source	aw trend	time trend	Assignment		Related functionality
1364	PLS, 3-WAY	↑	↑	water solvation shell [24], [25], [26]	ion hydrationproton hydrationwater vapour	Low firmness (mealiness) of apples (1364 nm, 1372 nm, 1382 nm) [71], wheat kernel hardness (1366 nm) [84], seed viability (1366 nm) [97]
1375	3-WAY			combination of symmetric and asymmetric stretch of H2O v1 + v3		
1382	3-WAY			water solvation shell [24], [27]		
1392	PCA, PLS, 3-WAY	~	↑	trapped water		Drying, dehydration, expulsion of cellular water, damage, stress, infection [39], [50], [89], [97], [98], [99], [100], [101]
1410	PLS, 3-WAY	↓	~	free water		Moisture content, water activity, seed viability [102], [103], [104]
1425	PCA, PLS, 3-WAY	↓	~	hydration band		Hydration of proteins, water activity, damage and defects [41], [76], [105], [106]
1436	PCA, PLS	~	↑	non-bonded O–H stretching first overtoneovertone of the OH bending mode of H5O2+		Phase transition, sugar-water interaction, hardness, seed viability [83], [84], [85], [97]
1455	PCA, PLS	↓	↓	water solvation shell (4, 5)		Water activity [53], viral infection in plants [107], damage [89]
1492	3-WAY	~ const	↓	water molecules with 4 hydrogen bonds		Damage/preservation (1496 nm) [51], [89]
1518	PLS, 3-WAY	↑	~	strongly bonded water		Preservation/ damage of materials, seed viability [98], [97], [99], [100], [105]

This work uncovered a very complex picture of water molecular structure in the stored rice germ, which, depending on the initial storage conditions – water activity and storage atmosphere is shaped differently and dictates modifications during storage.

The water molecular species which strongly characterized extreme levels of water activity levels were water solvation, proton hydration, strongly bound water and free water. Free water molecules are usually explained as the part of water that is able to participate in chemical reactions, the part which can act as a solvent, with the properties and reactivity more similar to the pure water [49] and which results in microbial spoilage when a_w greater than 0.6 [26], [50]. Massive decrease of free water molecules was found also to have an important role in preservation of living plant tissues upon extreme desiccation [51]. The strongly bound water is the water directly bound to other biological structures and cannot be easily removed from the food, for example by drying or squeezing; this part of water is not available for chemical reactions. The finding that samples stored at highest water activity level show highest absorbance of bound water and the lowest of free water, might be considered as unexpected. However, it should be noted that all levels of water activity in this study are below the 0.6 threshold, under which the food is considered stable towards microbial growth, however chemical and enzymatic reactions can occur even below this level.

Results reported in the present study indicate that – differently from what might have been expected – the water molecular species commonly referred to as free water (single water molecules, not involved in hydrogen bonds, and absorbing at 1410 nm) are not the water molecular structures that characterize high water activity for rice germ, under the conditions investigated. Throughout the study, during the entire observation period, and as documented in a previous study [45], the a_w levels of samples were monitored using a high-precision, high-accuracy instrument, and none of the samples showed any significant modifications in a_w, which testifies to the fact that single parameter like water activity is not descriptive enough to capture all the relevant factors contributing to the modifications during storage and that the phenomena of food modifications during storage is much more complex than

previously thought. This is also evident from the variety of water species that were found during aquaphotomics analysis and the observed complex dynamics of water molecular restructuring during storage. Further studies in this direction may help resolve the conflicting reports about the water activation energy because our results surely confirm the existence of highly complex water structure related to water activity and storage that cannot be explained by the simple model as suggested long time ago [26]. In addition to the bound water, the type of water which was associated with high water activity samples (and samples stored in air) was water represented by three wavelengths: 1364 nm (solvation shell with 1, 2 or 4 H_2O molecules), 1375 nm (proton hydrates) and 1382 nm (solvation shell with 1 or 4 H_2O molecules) [34]. This type of water is involved with hydration of charges. The entire spectral region from 1360 to 1385 nm was found to be rich in the absorbance bands attributed to hydration of small proton hydrates (up to 8 molecules in a water cluster): H^+(H2O) (1360 nm [52], 1369 nm [53], 1371 nm [54]), $H^+(H_2O)_3$ (1363 nm and 1380 nm [42], 1377 nm [42], [54]), $H^+(H_2O)_4$ (1359 nm [42], 1370 nm [42], [54], 1371 nm [43], [55], 1376 nm [56], 1385 nm and 1386 nm [42], [43]), $H^+(H_2O)_5$ (1381 nm [41], [54], 1371 nm [54]), $H^+(H_2O)_6$ (1370 nm [54], 1371 nm[56]), $H^+(H_2O)_7$ (1359 nm [54], [57], 1369.5 nm [56]), $H^+(H_2O)_8$ (1358 nm [57], 1369.5 nm [56], [57]), and at the end of this region larger clusters with more than 10 water molecules (1389 nm [54], [58]). As it was shown in aquagrams (Fig. 5) the absorbance at the entire region is higher for samples with highest water activity, so it is not possible to easily decompose and pinpoint the absorbance to only one particular water formation, nor it is likely the case. Rather, it seems to be a distribution of water species belonging to not-hydrogen bonded water with the higher-energy level states compared to the free water molecules (wavelengths lower than 1410 nm: lower wavelengths → higher frequency → higher energy). This indicates the possibility of joint influence of all three structures together – solvation shells around ions, proton hydrates and water vapour on proton mobility and proton transfer reactions.

A central place in proton transfer reactions is a dissociation step, occurring on a picosecond time scale, well after the rearrangement of hydrogen bonds [59]. The rate of proton dissociation was found to be decreasing with the increasing concentrations of salt in salt solutions, and to show a linear correlation with the water activity of

solutions, enabling estimation of a_w of electrolyte solutions based on the measurements of proton dissociation rate using fluorometric technique [60]. The dissociation rate parameter strongly depends on the water concentration and temperature, while proton transfer step itself, occurring in picoseconds, crucially depends on the solvent [58]. The proton transfer rate (mobility) was found to increase with the increasing temperature, while transfer kinetics is determined by availability of hydrating units (solvation water) in the surrounding aqueous media that can act as proton acceptors, as well as their specific structure (cluster) and size (critical size being 4 ± 1 water molecules in water cluster) [61].

Hence, our results suggest that water activity is a parameter describing the proton mobility within the food matrix. The measured single parameter, a_w, probably doesn't consider the influence of food matrix on the structure of water within. However, this water does interact with the rest of the aqueous media of food, creating channels and pores – a confined environment, thereby influencing the proton mobility (influence of space charge and boundary effects [39]). The interaction between the different food matrices and aqueous phase leads to different preservation/degradation occurrences during storage because the structure of food determines different possible proton transport mechanisms. This bound water takes into account this influence, and in our results, it is represented by the absorption at 1518 nm. In support of this observation, numerous studies have connected the absorbance of water species at this wavelength to the damage and preservation, as indicated in Table 2. It is important to emphasize here, that water activity by definition is the ratio of the water vapor pressure of the substance (food) to the water vapor pressure of pure water. At this point, we will temporarily leave out other solvation structures (solvated ions, solvated protons) from the discussion. The spectral results obtained here clearly connected the samples stored at higher a_w levels with the spectral pattern of water vapor, meaning - samples with higher water activity levels have more water in gaseous phase compared to the lower water activity samples. This gives rise to the higher vapor pressure which is an indication of evaporation rate, or the tendency of water to "escape" from the germ matrix (fugacity). The parameter a_w measured using dew point principal instruments is a measure of this tendency.

The water vapor spectral pattern also showed difference between the samples stored in the air from the samples stored in the modified atmosphere, where the air was removed (vacuum) or replaced by argon in order to prevent oxidation due to the oxygen from air. In both cases, modifying the atmosphere affected the evaporation rate of water. In the case of vacuum, the pressure inside the container was reduced through a vacuum pump, and the resulting pressure difference caused more rapid evaporation compared to the air conditions; this phenomenon is well-known and is used as a basic working principle of vacuum cooling techniques. On the other hand, studies have shown that not only reducing the pressure increases the diffusion coefficient of water vapor and consequently increases the evaporation rate, the nature of the ambient gas also has effects on diffusion coefficient and hence, the evaporation rate [63]. In fact, it was found that gases (except for helium and hydrogen) increase the evaporation rate of water, and out of eleven examined gases the largest rate was recorded for argon [64], [65]. This would mean that this part of water is simply eliminated from the samples with lower water activity, upon initial drying to that value and subsequently due to the influence of storage atmosphere. For the same water activity level, it seems that storage in vacuum or argon affects somewhat modifications in the first months of storage, but in an opposite way for samples stored at 0.36 and 0.45 a_w. However, while water vapour is most probably contributing to the observed absorbance pattern, it is important to take into account the role of hydrated protons, i.e. the connection between the water vapour evaporation/sorption and the proton mobility. As several studies found, the solvated protons display a marked preference for liquid/vapour interface i.e. the surface preference [66], [67], [68], [69], and even at low bulk water concentrations, the hydrated protons are the ones adsorbing at the water–air interface, showing substantially higher surface activity compared to that of the hydroxide ion, and strongly influencing the evaporation and condensation rates of water (reactivity, activity)[70].

Taking all these into account, we can summarize our observations, and propose the following interpretation. The spectral pattern at 1364 nm, 1375 nm, 1382 nm and 1518 nm, i.e. the higher absorbance at these wavelengths, being a dominant spectral feature of the sample with the highest water activity (a_w = 0.55), describes the proton mobility within the food matrix, taking into account the influence of food matrix itself on

the rate of dissociation and proton transfer, and more accurately describes the possibility of chemical reactions, or “real” water activity, compared to a single number measured by current measurement devices. On the other hand, the samples stored at lower activity levels, and in the modified atmosphere are strongly characterized by the following four water absorbance bands – 1410 nm, 1425 nm, 1455 nm and 1474 nm. In a previous aquaphotomics study, which aimed to relate water molecular structure with the textural properties of the apples, the high values at the water vapour bands were found to be a common characteristic of a mealy (porous, dry) texture, while the crispy and juicy texture of the apples was characterized by high absorbance in the region of free, medium and hydrogen bonded water (1410 – 1492 nm) [71]. The similarity between the patterns found in that work and in our study, implies the existence of differences in the rice germ from the aspect of texture. It is well-known that texture of the food depends on the mechanical and structural properties of the food, and that it can change during the storage or depending on the storage conditions [72]. The uptake or release of water to the environment, as well as mechanical properties of the samples are dependent on the macromolecular interactions within the sample matrix. When the manipulation of water activity before storage is performed, by removal or binding of the water, it also alters these properties, resulting in different internal space and levels of water confinement. The texture is closely related to the structure of the sample tissue [73]. The differences in texture, in this work, were revealed based on differences in water spectral pattern of three rice germ samples. The spectral pattern of high water activity samples showed strongly bound water and water vapor features in the respective water spectral pattern. These water molecular species indicated that while the sample appears solid (preserved tightly bound water to the undamaged cell wells), there is more space between the individual cells filled with the air and water vapor. As opposed to this, the samples stored at lower water activity levels, show preservation of cells packed with liquid with more variations in water species, as evidenced by the spectral pattern of free water (1410 nm), protein hydration water (1425 nm), bulk water (1455 nm) and water molecules with three hydrogen bonds (1474 nm), all of which can be released upon cell rupture producing the “juicy” feeling. In a study which examined the water vapour diffusivity in vitreous and mealy wheat endosperm, it was found that mealy endosperm due to the

intracellular air spaces allowed much higher water vapor diffusivity (from 1.8 to 4.6 times higher), compared to the vitreous endosperm which has extremely small air spaces, if any, and hence much higher density [74]. Together, these results indicate that the higher water activity in rice germ samples is associated with the texture which allows easier diffusion of water vapor through the sample matrix, i.e. higher rate of moisture absorption/desorption (similar to the sponge), while samples stored at a lower water activity may have undergone glassy transition, which resulted in preservation of biological structures and the water species associated.

The differences in texture properties would result in different behavior during processing of the rice germ, which is very important from the aspect of predicting particular behavior and outcomes depending on the initial water activity and storage conditions. However, while it is well-recognized that water activity affects the textural properties of the food [75], the relationship between the water structure and resulting textural properties of food is not clear, and our study showed that spectral pattern of water can be used as a descriptor of the texture, providing the clues to which type of processing (frying, baking etc.) would be a better choice. Further, it clearly showed that the water activity is related not to the simply called “free water”, but to the richness of energy states of water which provides diffusivity/mobility of water within the food matrix and consequently exchange with the storage environment (moisture desorption/absorption). These processes showed a clear relation to the textural properties.

The last two absorbance bands which are found to be important for water activity are 1425 nm and 1455 nm. These bands can be assigned to intermediate water structures (not free, not interconnected water molecules), specifically 1425 nm is associated with hydration water [34], [40], [76], while 1455 nm is ascribable to so called, bulk water or also adsorbed water [77] (physically adsorbed at the pores of the food microstructure). Previous works reported that interplay between these two bands is strongly related to the water activity [76]. Whether the intensity of the bands at 1425 nm and 1455 nm is going to change in the same or opposite direction upon the increase/decrease of water content seems to be different for different systems [76], [78]. In another words, it is dependent on the nature of the sample matrix (the pore sizes of the sample matrix, the evaporation rate during drying etc. [79]). Over time, intensity at 1455 nm somewhat

decreases, showing decrease in bulk water, while a less marked behaviour is detectable for intensity at 1425 nm. There is another possible interpretation of these two bands. The band 1425 nm is very close to the band 1428 nm reported to increase in the intensity upon drying of agricultural materials, which may be related to the glucose molecules, the basis of starch and cellulose in those materials, which can become more visible upon drying [80]. The band at 1455 nm can also be assigned to the first overtone of O–H stretching of histidine that is a well-known degradation product developed during long storage of food products containing proteins, such as rice germ [80]. Another interesting finding is that irrespective of water activity levels, when the samples' water matrix changes over time, the intensity of band at 1492 nm changes in the same way for all the samples. This band can be assigned to water molecules with 4 hydrogen bonds [57]. The overtone of this band (located around 996 nm) was proposed to be due to the OH stretching of water interacting with protein or OH stretching vibration of proteins [82]. Either way, for our results, this indicates that in the rice germ during storage, some proteins undergo changes in a very similar way, despite initial modulation of water activity in the samples – the water activity does not affect this feature of rice germ.

The changes in rice germ also showed restructuring of water matrix as a function of storage time, also. In fact, the difference between samples with different water activity was limited to specific regions of the spectra, i.e. only involved certain water molecular species. The absorbance bands which showed the largest changes were located at 1436 nm and 1392 nm. The contribution of absorbance at 1436 nm wavelength increases with time. This band was reported to be important for transition of amorphous to crystalline states of sugars – lactose [61], raffinose and sucrose [84]. In the latter study, 1436 nm absorption was assigned to non-bonded O–H stretching first overtone band, which had different shape for amorphous samples compared to the crystalline samples. A very close band, located at 1440 nm and ascribed to a water dimer, was found of critical importance for preservation of plant tissues upon extreme desiccation. The plants which can survive desiccation, showed massive accumulation of water dimers, which was connected to the influence of reducing sugars – sucrose and raffinose on the formation of glassy, vitreous state [51]. Another literature source reports close band (1430 nm) to increase in intensity upon drying of agricultural

materials, revealing carbohydrate units of starch and cellulose upon drying [80]. Further, this exact band was found to be positively correlated with wheat kernel hardness in a study of wheat kernel water extracts [85]. All these studies suggest that, most probably this is not a carbohydrate band, but water-carbohydrate interaction band related to the phase transition of water. Since it is very well known that water works as an effective plasticizer in food matrices, decreasing glass transition temperature and mechanical resistance [86], taking together the interpretations of mentioned studies, we can suggest that 1436 nm water band in our results is indicative of changes in plasticity/hardness of the germ matrix happening during storage as a result of phase transition of water. Considering that rice germs are rich in carbohydrates, it can be deduced that, during storage time the process of phase transition happens which leads to changes in hardness of the germ. The monosaccharides, in fact, have a glass transition in the vicinity of room temperature, which can be a relevant aspect in rice germ modifications along the time, when the product is microbially stable due to the low a_w level [87]. When the regression vectors of PLSR analysis using time as the dependent variable are examined (Fig. 4d), the highest coefficient at the 1436 nm can be found for the samples stored at the lowest water activity level, suggesting that they are prone to the strongest changes during time with the respect to the phase transition and hardness.

The band at 1392 nm is a trapped water band [39] and it shows positive correlation with the time of storage, indicating that water molecules become more confined upon the depletion of bulk and free water. This band is found to be strongly associated with dehydration in several works [88], [89]. This points out the process of dehydration over time. The importance of water vapor bands for storage process (1364, 1375 and 1382 nm) supports this interpretation, suggesting evaporation of water from the rice germ takes place.

Conclusion and scientific impacts

In the light of the results presented, it is possible to outline some comprehensive conclusions. Samples stored at higher a_w level are characterized by high intensities of NIR spectral bands associated with strongly bonded water, solvation water, proton

hydrates and water vapor, and lower quantity of free, hydration and bulk water. The same is true for the samples stored in air, compared to the samples stored in vacuum and argon. The higher water activity was strongly associated with the presence of water vapor phase and highly surface-active proton hydrates, while this part of water was already removed from the rice germ samples stored at lower water activity owing to the drying. Similarly, the storage in vacuum and argon caused higher evaporation rate and loss of this water type. The water activity, hence, was strongly related to the presence of water in gaseous phase and proton hydrates at the surface, *i.e.* the water species which provide fast proton mobility water, in other words, species which with easiness can leave the surface or diffuse through the sample matrix, and also to interact with the environment through sorption.

The levels of initial water activity and storage atmosphere also resulted in different textural properties, where soft and mealy texture was associated with high water activity and storage in air, while lower water activity and storage in modified atmosphere resulted in restructuring water in rice germ towards amorphous state, in which the hydration of the biological structures, and much more variety of the water molecular structure was preserved in the cells, including free, bulk and hydrogen bonded water.

While it may seem surprising, it must be emphasized that the range of water activity we examined was between 0.36 and 0.55 and, under these conditions, deterioration and microbial spoilage cannot occur. The water activity within this range was associated with the mobility of water/easiness of water to move within the food matrix and evaporate, as witnessed by monitoring changes over time. On the other hand, this probably also means that samples with higher water activity, if stored under atmosphere with higher relative humidity, would undergo an opposite process – they would absorb the moisture from the air more easily. In terms of storage atmosphere, storage in air has the same effects.

Regarding the modifications along time, they involve restructuring of different water molecular conformations to water vapour, solvation water and proton hydrates (1364 nm, 1375 nm and 1382 nm), in the samples stored at lower water activity, while in the samples with high water activity it happens directly. The restructuring and

evaporation are producing changes at the trapped water/dehydration band (1392 nm) and changes at the carbohydrate-water interaction band (1436 nm) related to the hardness of the germ.

Taken together, these results explain that storage of rice germ with respect to water activity can be described as a process of dehydration and decrease of bulk water, free water and changes in hydration of biological structures, then evaporation and changes in hardness. All these processes are dependent upon the initial conditions – initial water activity and storage atmosphere, which determine how fast the dynamics of water restructuring can occur. These changes, while not affecting the microbial stability, do affect physical properties such as structure and texture, which consequently have substantial influence on consumers' perception of quality and food processing. While today it is generally accepted that water activity is more closely related to the microbial, chemical and physical properties of the food and not its mere moisture content [75], this study for the first time revealed that what defines water activity and governs how the food will be modified during storage is the water molecular structure of the specific food matrix. While the water activity is one number, a single parameter related to only water vapour, this study revealed far more complex picture, showing very detailed, fine water molecular structure dynamics and how it is related to the expressed properties of the samples as well as their modifications depending on the storage atmosphere and initial water activity. This study shows that there is more complexity, even compared to the studies of water in food using NMR that detects only three types of water (structural, multilayer-surface adsorbed water and bulk water) [90]. Understanding the role of water molecular structure dynamics behind the water activity is demonstrated to be a fundamental step for explaining the nature of the processes that occur during storage in one complex biological matrix – rice germ. The studies such as this one, performed on other biological matrices can help elucidate the complex relationship between the degradation or preservation and water activity and also help explaining the differences in degradation of materials stored at the same water activity but achieved through different processes – desorption and adsorption [91] and thus bridge the gap between the fundamental research and practical applications. Monitoring the functionality of water molecular formations during storage also provided once more the evidence for the

aquaphotomics concept of water being a molecular mirror and an immense source of information. And it may in the future provide the predictions of shelf-life depending on the storage conditions based on non-destructive aquaphotomics near infrared spectral monitoring, something which is still unresolved problem in science [26].

[1] Siesler, H. W., Kawata, S., Heise, H. M., & Ozaki, Y. (Eds.). (2008). Near-infrared spectroscopy: principles, instruments, applications. John Wiley & Sons.

[2] Norris, K. H. (1996). History of NIR. Journal of Near Infrared Spectroscopy, 4(1), 31-37.

[3] Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., & De La Pezuela, C. (1998). Near-infrared spectroscopy in the pharmaceutical industry. Critical review. Analyst, 123(8), 135R-150R.

[4] Shenk, J. S., Workman Jr, J. J., & Westerhaus, M. O. (2007). Application of NIR spectroscopy to agricultural products. In Handbook of near-infrared analysis (pp. 365-404). CRC Press.

[5] Martin, P. A. (2002). Near-infrared diode laser spectroscopy in chemical process and environmental air monitoring. Chemical Society Reviews, 31(4), 201-210.

[6] Porep, J. U., Kammerer, D. R., & Carle, R. (2015). On-line application of near infrared (NIR) spectroscopy in food production. Trends in Food Science & Technology, 46(2), 211-230.

[7] Trygg, J., & Wold, S. (1998). PLS regression on wavelet compressed NIR spectra. Chemometrics and Intelligent Laboratory Systems, 42(1-2), 209-220.

[8] Bilic-Zulle, L. (2011). Comparison of methods: Passing and Bablok regression. Biochemia medica, 21(1), 49-52.

[9] Tsenkova, R. (2009). Aquaphotomics: dynamic spectroscopy of aqueous and biological systems describes peculiarities of water. Journal of Near Infrared Spectroscopy, 17(6), 303-313.

[10] International Olive Oil Council. (2003). Trade standard applying to olive oil and olive-pomace oil. Resolution No. Res-3/89-Iv/03, COI/T. 15/NC No. 3/Rev. 1.

[11] Forina, M., Armanino, C., Lanteri, S., & Tiscornia, E. (1983). Classification of olive oils from their fatty acid composition. In Food research and data analysis: proceedings from the IUFoST Symposium, September 20-23, 1982, Oslo, Norway/edited by H. Martens and H. Russwurm, Jr. London: Applied Science Publishers, 1983.

[12] Schwingshackl, L., & Hoffmann, G. (2014). Monounsaturated fatty acids, olive oil and health status: a systematic review and meta-analysis of cohort studies. Lipids in health and disease, 13(1), 1-15.

[13] Casale, M., Oliveri, P., Casolino, C., Sinelli, N., Zunin, P., Armanino, C., ... & Lanteri, S. (2012). Characterisation of PDO olive oil Chianti Classico by non-selective (UV–visible, NIR and MIR spectroscopy) and selective (fatty acid composition) analytical techniques. Analytica Chimica Acta, 712, 56-63.

[14] Blanco, M., & Villarroya, I. N. I. R. (2002). NIR spectroscopy: a rapid-response analytical tool. TrAC Trends in Analytical Chemistry, 21(4), 240-250.

[15] Mailer, R. J. (2004). Rapid evaluation of olive oil quality by NIR reflectance spectroscopy. Journal of the American Oil Chemists' Society, 81(9), 823-827.

[16] Manley, M., & Eberle, K. (2006). Comparison of Fourier transform near infrared spectroscopy partial least square regression models for South African extra virgin olive oil using spectra collected on two spectrophotometers at different resolutions and path lengths. Journal of near infrared spectroscopy, 14(2), 111-126.

[17] Wesley, I. J., Pacheco, F., & McGill, A. E. J. (1996). Identification of adulterants in olive oils. Journal of the American Oil Chemists' Society, 73(4), 515-518.

[18] Downey G, McIntyre P, Davies AN. Geographic classification of extra virgin olive oils from the eastern Mediterranean by chemometric analysis of visible and near-infrared spectroscopic data. Appl Spectrosc. 2003 Feb;57(2):158-63.

[19] Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), 109-130.

[20] Oliveri, P., Malegori, C., Simonetti, R., & Casale, M. (2019). The impact of signal pre-processing on the final interpretation of analytical outcomes–A tutorial. Analytica chimica acta, 1058, 9-17.

[21] Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. Technometrics, 11(1), 137-148.

[22] Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent laboratory systems, 44(1-2), 175-185.

[23] Passing, H., & Bablok, W. (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I.

[24] L.N. Bell, T.P. Labuza, Practical Aspects of Moisture Sorption Isotherm Measurement and Use, 2nd Ed., AACC Eagan, MN, 2000.

[25] Y.H. Roos, in: Water activity and plasticization in Food Shelf-Life Stability: Chemical, Biochemical, and Microbiological, CRC Press, 2001, pp. 3–36.

[26] T.P. Labuza, The effect of water activity on reaction kinetics of food deterioratio, Food Technol. 34 (1980) 36–41.

[27] N. Campbell, G. Campbell Scott, A. Campbell, R.O. Fontana, Water activity determination using near-infrared spectroscopy. 16 (2004).

[28] A. Mason, A. Al-Shammaa, O. Alvseike, Apparatus and method for measuring water activity in food products 30 (2017).

[29] T.P. Labuza, Properties of water as related to the keeping quality of foods, in: Proceedings of the Third International Congress of Food Science & Technology in Proceedings of the Third International Congress of Food Science & Technology, 1970, pp. 618–635.

[30] F. Shahidi, Y. Zhong, Lipid oxidation and improving the oxidative stability, Chem. Soc. Rev. 39 (2010) 4067–4079.

[31] A. Moongngarm, N. Daomukda, S. Khumpika, Chemical Compositions, Phytochemicals, and Antioxidant Capacity of Rice Bran, Rice Bran Layer, and Rice Germ. APCBEE Procedia (2012) https:/doi.org/10.1016/j. apcbee.2012.06.014

[32] P.D. Babu, P.D. Babu, R.S. Subhasree, R. Bhakyaraj, R. Vidhyalakshmi, Brown Rice-Beyond the Color Reviving a Lost Health Food- A Review, Am. J. Agron. 2 (2009) 67–72.

[33] M. Mathlouthi, Water content, water activity, water structure and the stability of foodstuffs, Food Control 12 (2001) 409–417.

[34] R. Tsenkova, Aquaphotomics: Dynamic spectroscopy of aqueous and biological systems describes peculiarities of water, J. Near Infrared Spectrosc. 17 (2009) 303–313.

[35] J. Munan, R. Tsenkova, AquaphotomicsFrom Innovative Knowledge to Integrative Platform in Science and Technology, Molecules 24 (2019) 2742.

[36] R. Tsenkova, J. Munan, B. Pollner, Z. Kovacs, Essentials of aquaphotomics and its chemometrics approaches, Front. Chem. 6 (2018) 363.

[37] A. Putra, F. Faridah, E. Inokuma, R. Santo, Robust spectral model for low metal concentration measurement in aqueous solution reveals the importance of water absorbance bands, J. Sains dan Teknol. Reaksi 8 (2010).

[38] R. Tsenkova, Aquaphotomics: Water in the biological and aqueous world scrutinised with invisible light, Spectrosc. Eur. 22 (2010) 6–10.

[39] D. Koji, R. Tsenkova, K. Tomobe, K. Yasuoka, M. Yasui, Water confined in the local field of ions, ChemPhysChem 15 (2014) 4077–4086.

[40] E. Chatani, Y. Tsuchisaka, Y. Masuda, R. Tsenkova, Water molecular system dynamics associated with amyloidogenic nucleation as revealed by real time near infrared spectroscopy and aquaphotomics, PLoS One 9 (2014) e101997.

[41] H.A. Schwarz, Gas phase infrared spectra of oxonium hydrate ions from 2 to 5 , J. Chem. Phys. 67 (1977) 5525–5534.

[42] M. Okumura, L.I. Yeh, J.D. Myers, Y.T. Lee, Infrared spectra of the solvated hydronium ion: Vibrational predissociation spectroscopy of mass-selected H3O+ñ(H2O)nñ(H 2)m, J. Phys. Chem. 94 (1990) 3416–3427.

[43] L.I. Yeh, M. Okumura, J.D. Myers, J.M. Price, Y.T. Lee, Vibrational spectroscopy of the hydrated hydronium cluster ions H 3O+ñ(H2O)n (n= 1, 2, 3), J. Chem. Phys. 91 (1989) 7319–7330.

[44] R. Tsenkova, Z. Kovacs, Y. Kubota, Aquaphotomics: Near infrared spectroscopy and water states in biological systems in Membrane Hydration, E. A. DiSalvo, Ed. (Springer, Cham, 2015), pp. 189211

[45] C. Malegori et al., A modified mid-level data fusion approach on electronic nose and FT-NIR data for evaluating the effect of different storage conditions on rice germ shelf life, Talanta 206 (2020).

[46] F. Dahms et al., The Hydrated Excess Proton in the Zundel Cation H2 + : The evaporation rate, and water absorption band depths in SWIR reflectance Role of Ultrafast Solvent Fluctuations, Angew. Chemie Int. Ed. 55 (2016) 10600–10605.

[47] W. Kulig, N. Agmon, A clusters-in-liquid method for calculating infrared spectra identifies the proton-transfer mode in acidic aqueous solutions, Nat. Chem. 5 (2013) 29–35.

[48] M. Leuchs, G. Zundel, easily polarizable hydrogen bonds in aqueous solutions of acids. Perchloric acid and trifluoromethane sulphonic acid, J. Chem. Soc. Faraday Trans. 2 Mol. Chem. Phys. 74 (1978) 2256–2267.

[49] P.S. Taoukis, P.S. Taoukis, T.P. Labuza, I.S. Saguy, in: Kinetics of food deterioration and shelf-life prediction in Handbook of Food Engineering Practice, CRC Press, 1997, pp. 361–403.

[50] A.A. Gowen, Water and Food Quality, Contemp. Mater. 1 (2012) 31–37.

[51] S. Kuroki et al., Water molecular structure underpins extreme desiccation tolerance of the resurrection plant Haberlea rhodopensis, Sci. Rep. 9 (2019) 3049.

[52] J.C. Jiang et al., Infrared spectra of H+(H2O)58 clusters: Evidence for symmetric proton hydration, J. Am. Chem. Soc. 122 (2000) 1398–1410.

[53] E.G. Diken et al., Fundamental excitations of the shared proton in the H 3O 2- and H 5O 2+ complexes, J. Phys. Chem. A 109 (2005) 1487–1490.

[54] J.M. Headrick et al., Spectral signatures of hydrated proton vibrations in water clusters, Science (80-.) 308 (2005) 1765–1769.

[55] G.E. Douberly, R.S. Walters, J. Cui, K.D. Jordan, M.A. Duncan, Infrared spectroscopy of small, protonated water clusters, H +(H2O) n (n = 25): Isomers, argon tagging, and deuteration, J. Phys. Chem. A 114 (2010) 4570– 4579.

[56] K. Mizuse, A. Fujii, Tuning of the Internal Energy and Isomer Distribution in Small Protonated Water Clusters H + (H 2 O) 48: An Application of the Inert Gas Messenger Technique, J. Phys. Chem. A 116 (2012) 4868–4877.

[57] D. Wei, D.R. Salahub, Hydrated proton clusters: Ab initio molecular dynamics simulation and simulated annealing, J. Chem. Phys. 106 (1997) 6086–6094.

[58] J.W. Shin et al., Infrared signature of structures associated with the H+ (H2O) n (n= 6 to 27) clusters, Science. 304 (5674) (2004 May 21) 1137–1140.

[59] N. Agmon, Elementary steps in excited-state proton transfer, J. Phys. Chem. A 109 (2005) 13–35.

[60] D. Huppert, E. Kolodney, M. Gutman, E. Nachliel, Effect of water activity on the rate of proton dissociation, J. American Chemical Society. 104 (25) (1982) 6949–6953.

[61] J. Lee, R.D. Griffin, G.W. Robinson, 2-Naphthol: A simple example of proton transfer effected by water structure, J. Chem. Phys. 82 (1985) 4920–4925.

[62] P. Knauth, E. Sgreccia, A. Donnadio, M. Casciola, M.L. Di Vona, Water Activity Coefficient and Proton Mobility in Hydrated Acidic Polymers, J. Electrochem. Soc. 158 (2011) B159.

[63] K. Sefiane, S.K. Wilson, S. David, G.J. Dunn, B.R. Duffy, On the effect of the atmosphere on the evaporation of sessile droplets of water, Phys. Fluids 21 (2009) 062101.

[64] K.H. Kingdon, Enhancement of the evaporation of water by foreign molecules adsorbed on the surface, J. Phys. Chem. 67 (1963) 2732.

[65] W.W. Mansfield, Influence of gases on the rate of evaporation of water, Nature 205 (1965) 278.

[66] M.K. Petersen, S.S. Iyengar, T.J.F. Day, G.A. Voth, The hydrated proton at the water liquid/vapor interface, J. Phys. Chem. B 108 (2004) 14804–14806.

[67] C. Rad≈ge, V. Pflumio, Y.R. Shen, Surface vibrational spectroscopy of sulfuric acid-water mixtures at the liquid-vapor interface, Chem. Phys. Lett. 274 (1997) 140–144.

[68] P.B. Miranda, Y.R. Shen, Liquid interfaces: A study by sum-frequency vibrational spectroscopy, J. Phys. Chem. B 103 (1999) 3292–3307.

[69] S. Das, M. Bonn, E.H.G. Backus, The Surface Activity of the Hydrated Proton Is Substantially Higher than That of the Hydroxide Ion, Angew. Chemie Int. Ed. 58 (2019) 15636–15639.

[70] A.M. Rizzuto, E.S. Cheng, R.K. Lam, R.J. Saykally, Surprising Effects of Hydrochloric Acid on the Water Evaporation Coefficient Observed by Raman Thermometry, J. Phys. Chem. C 121 (2017) 4420–4425.

[71] M. Vanoli et al., Water spectral pattern as a marker for studying apple sensory texture, Adv. Hortic. Sci. 32 (2018) 343–351.

[72] S.G. Gwanpua et al., Pectin modifications and the role of pectin-degrading enzymes during postharvest softening of Jonagold apples, Food Chem. 158 (2014) 283–291.

[73] A. Arefi et al., Mealiness Detection in Agricultural Crops: Destructive and Nondestructive Tests: A Review, Compr. Rev. Food Sci. Food Saf. 14 (2015) 657–680.

[74] G.M. Glenn, R.K. Johnston, Water vapor diffusivity in vitreous and mealy wheat endosperm, J. Cereal Sci. 20 (1994) 275–282.

[75] J. Chirife, A. J. Fontana Jr., Introduction: Historical ighlights of water activity research in Water Activity in Foods, 1st Ed., G. V. Barbosa-Canovas, A. J. Fontana Jr., S. Schmidt, T. P. Labuza, Eds. (Blackwell Publishing and the Institute of Food Technologists, 2007), pp. 314.

[76] A. Heiman, S. Licht, Fundamental baseline variations in aqueous near-infrared analysis, Anal. Chim. Acta 394 (1999) 135–147.

[77] J. Ma, B. Zhou, H. Zhang, W. Zhang, Z. Wang, activated municipal wasted sludge biochar supported by nanoscale Fe/Cu composites for tetracycline removal from water, Chem. Eng. Res. Des. 149 (2019) 209–219.

[78] Y. Ozaki, T. Miura, K. Sakurai, T. Matsunaga, Nondestructive Analysis of Water Structure and Content in Animal Tissues by FT-NIR Spectroscopy with Light- Fiber Optics Part I: Human Hair, Appl. Spectrosc. 46 (1992) 875–878.

[79] J. Tian, W.D. Philpot, Relationship between surface soil water content, spectra, Remote Sens. Environ. 169 (2015) 280–289.

[80] P. Williams, Influence of water on prediction of composition and quality factors: The Aquaphotomics of low moisture agricultural materials, J. Near Infrared Spectrosc. 17 (2009) 315–328.

[81] R. Singh Yadav, Studies of Histidine, Phenylalanine Complexes of Oxovanadium (IV) Derived from Acetylacetone, Nat. Preced. (2010), https:// doi.org/10.1038/npre.2010.4378.1.

[82] S. ai, Y. Ozaki, Short-wave near-infrared spectroscopy of biological fluids. 1. Quantitative analysis of fat, protein, and lactose in raw milk by partial least- squares regression and band assignment, Anal. Chem. 73 (2001) 64–71.

[83] R.A. Lane, G. Buckton, The novel combination of dynamic vapour sorption gravimetric analysis and near infra-red spectroscopy as a hyphenated technique, Int. J. Pharm. 207 (2000) 49–56.

[84] P.E. Luner, J.J. Seyer, Assessment of crystallinity in processed sucrose by near- infrared spectroscopy and application to lyophiles, J. Pharm. Sci. 103 (2014) 2884–2895.

[85] B.H. Hong, G.L. Rubenthaler, R.E. Allan, Wheat pentosans. II. Estimating kernel hardness and pentosans in water extracts by near-infrared reflectance, Cereal Chem. 66 (5) (1989) 374–377.

[86] P. Pittia, G. Sacchetti, Antiplasticization effect of water in amorphous foods, A review. Food Chem. 106 (2008) 1417–1427.

[87] Y. H. Roos, Water Activity and Glass Transition in Water Activity in Foods, G. Barbosa-Canovas, A. Fontana Jr, S. J. Schmidt, T. P. Labuza, Eds. (Blackwell Publishing Ltd, 2007), pp. 2945.

[88] T.M.P. Cattaneo, G. Cabassi, M. Profaizer, R. Giangiacomo, Contribution of light scattering to near infrared absorption in milk, J. Near Infrared Spectrosc. 17 (2009) 337–343.

[89] A.A. Gowen, C. Esquerre, C.P. ODonnell, G. Downey, R. Tsenkova, Use of near infrared hyperspectral imaging to identify water matrix co-ordinates in mushrooms (Agaricus bisporus) subjected to mechanical vibration, J. Near Infrared Spectrosc. 17 (2009) 363–371.

[90] S. J. Schmidt, Water mobility in foods in Water Activity in Foods - Fundamentals and Applications, G. V. Barbosa-Canovas, A. . Fontana Jr, S. Schmidt, T. P. Labuza, Eds. (Blackwell Publishing and the Institute of Food Technologists, 2007), pp. 47108

[91] S.R. Delwiche, R.E. Pitt, K.H. Norris, Sensitivity of Near-Infrared Absorption to Moisture Content Versus Water Activity in Starch and Cellulose, Cereal Chem. 69 (1992) 107–109.

[92] P. Oliveri, C. Malegori, R. Simonetti, M. Casale, Analytica Chimica Acta The impact of signal pre-processing on the fi nal interpretation of analytical outcomes e A tutorial, Anal. Chim. Acta 1058 (2019) 9–17.

[93] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.

[94] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta (1986), https://doi.org/10.1016/0003-2670(86)80028-9.

[95] A.K. Smilde, Three-way analyses problems and prospects, Chemom. Intell. Lab. Syst. 15 (1992) 143–157.

[96] P.M. Kroonenberg, Three-mode principal component analysis. theory and applications DSWO, Press, 1983.

[97] L.M. Kandpal, S. Lohumi, M.S. Kim, J.S. Kang, B.K. Cho, Near-infrared hyperspectral imaging system coupled with multivariate methods to predict viability and vigor in muskmelon seeds, Sensors Actuators, B Chem. 229 (2016) 534–544.

[98] R. Giangiacomo, P. Pani, S. Barzaghi, Sugars as a Perturbation of the Water Matrix, J. Near Infrared Spectrosc. 17 (2009) 329–335.

[99] C. Esquerre, A.A. Gowen, C.P. ODonnell, G. Downey, Initial Studies on the Quantitation of Bruise Damage and Freshness in Mushrooms Using Visible- Near-Infrared Spectroscopy, J. Agric. Food Chem. 57 (2009) 1903–1907.

[100] M. Tigabu, P.C. OdÕn, Discrimination of viable and empty seeds of Pinus patula Schiede & Deppe with near-infrared spectroscopy, New For. 25 (2003) 163– 176.

[101] W. Wang, et al., Near-infrared hyperspectral reflectance imaging for early detection of sour skin disease in Vidalia sweet onions. Am. Soc. Agric. Biol. Eng. Annu. Int. Meet. 2010, ASABE 2010 4, 34153436 (2010).

[102] X. He et al., Rapid and nondestructive measurement of rice seed vitality of different years using near-infrared hyperspectral imaging, Molecules 24 (2019) 2227.

[103] E. Achata et al., A study on the application of near infrared hyperspectral chemical imaging for monitoring moisture content and water activity in low moisture systems, Molecules 20 (2015) 2611–2621.

[104] K. Phetpan, V. Udompetaikul, P. Sirisomboon, In-line near infrared spectroscopy for the prediction of moisture content in the tapioca starch drying process, Powder Technol. 345 (2019) 608–615.

[105] J. akota Rosi, J. Munan, I. Mileusni, B. Kosi, L. Matija, Detection of protein deposits using NIR spectroscopy, Soft Mater. 14 (2016) 264–271.

[106] B. Chu et al., Development of noninvasive classification methods for different roasting degrees of coffee beans using hyperspectral imaging, Sensors 18 (2018) 1259.

[107] R. A. Naidu, E.M. Perry, F.J. Pierce, T. Mekuria, The potential of spectral reflectance technique for the detection of Grapevine leafroll-associated virus-3 in two red-berried wine grape cultivars, Comput. Electron. Agric. 66 (2009) 38–45.

CHAPTER 3: NIR SPECTROSCOPY, PORTABLE DEVICE APPLICATIONS

In the last decades, near-infrared (NIR) spectroscopy (800–2500 nm) proved to be one of the most efficient analytical methods for quality control and process monitoring in the pharmaceutical and food fields. The increasing interest in developing NIR-based applications has led to a rapid technical progress in terms of instruments [1]. In more detail, the need to have flexible devices that ensure good analytical performances directly for field measurements or for the monitoring of complex manufacturing processes allowed to move from traditional benchtop spectrophotometers to more and more portable ones, until reaching miniaturized NIR sensors [2]

Thanks to a long-term collaboration with Viavi Solutions, an American company leader in designing and producing portable NIR devices (MicroNIR), I had the possibility to develop ad-hoc solutions based on the implementation of MicroNIR sensors and chemometrics for two industrial case-studies with the aim to define strategies for process monitoring.

Nowadays, NIRS is in fact considered one of the most powerful Process Analytical Technologies for the real-time process monitoring in manufacturing companies, especially in the pharmaceutical industry. PAT was formally introduced through the FDA guidance document in 2004, which allowed to build the new science-oriented and risk-based approach for quality control, together with the cGMP [3]. Thanks to the online implementation of a PAT system, it is possible to obtain a global understanding about the process with the aim of achieving a predefined quality product, even before testing it. A process can be considered well understood if all the variability sources are well known, well controlled, and if the quality attributes of the product can be efficiently predicted over the design space.

This third chapter deals with two industrial projects in which a miniaturized NIR sensor has been tested for the monitoring of blending processes in pharmaceutical and food

fields, respectively. Blending is a critical unit operation in manufacturing, as it is a prerequisite for obtaining the homogeneous distribution of mixture components. NIR spectra, properly acquired along the mixing process, can be used for the detection of the endpoint of the blending and for performing a continuous real-time verification of the process [4]. In the first paragraph of this chapter, a chemometric qualitative strategy for assessing the homogeneity of the powder blending of a zootechnical formulation has been reported. The present approach was based on the alternative application of the Moving Block Standard Deviation (MBSD) [5] method and the following development of a Multivariate Statistical Process Control (MSPC) model [6] based on Principal Component Analysis for building multivariate control charts. MBSD is one of the most applied methods for endpoint detection in blending processes; it works summarizing the variance in contiguous blocks of spectra for evaluating the variability of the blend over time. In this case, MBSD has been used in the direction of the samples for obtaining Standard Deviation (SD) spectra, to minimize the contribution of systematic effects. On the SD spectra, an MSPC model has been calculated for defining the space of information related to endpoint observations. Consequently, two MSPC control charts have been built, in which observations of new batches are represented [6]:

- Hoteling's T^2 chart, which allows to measure the in-model variation of the sample, for evaluating the stability of the process over time.

- Q-statistics (Q) chart, which represents the variance unaccounted for by the model, for testing whether any perturbations of the system break the correlations observed for the endpoint observations.

A multi-step qualitative approach for the real-time blending monitoring of a food powder formulation has been reported in the second paragraph of this chapter. Qualitative approaches have the advantages of not requiring mechanical external sampling and reference analysis for calculating the calibration model (typical for quantitative approaches). In this project, three different qualitative chemometric strategies have been tested in order to define pros and cons of each technique. In more detail, a critical comparison among MBSD, a PCA-MSPC based model and

Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) [7] have been proposed. MCR-ALS is a powerful tool to extract qualitative and quantitative information from a set of NIR spectra. This approach can provide the concentration profiles and pure spectral fingerprints for all compounds involved in the mixing by using only the spectral information acquired along the process.

MCR-ALS assumes that the original set of process observations behaves following a bilinear model, which is the multiwavelength extension of Lambert-Beer's law and is described by the following expression[8]:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E}$$

where $\mathbf{X}$ is the matrix of original spectra collected for a single or multiple batches. $\mathbf{S}^{\mathbf{T}}$ contains the pure spectra signatures of the components needed to describe the process and $\mathbf{C}$ the related concentration profiles. $\mathbf{E}$ is the matrix with the residual information not explained by the model, related to the experimental error.

Chapter 3

3.1 A MOVING BLOCK- PCA BASED APPROACH FOR THE NIR REAL – TIME MONITORING AND VERIFICATION OF A BLENDING PROCESS.

Scientific background and aim of the work

The concept of quality by design (QbD), adopted in the pharmaceutical field through recent quality regulatory initiatives such as FDA's Process Analytical Technology (PAT) [8]. Initiative, ICH Guidance Q8 [9] and Q9 [10], is based on the identification of predefined objectives with the aim to improve the product and process understanding. The application of QbD starts with the definition of the requirements of the final product: pharmaceutical form, delivery system, pharmacodynamic and pharmacokinetic properties [11]. In this context, to obtain the desired finished product, it is crucial to evaluate the critical characteristics of the raw materials (active ingredients, excipients, process materials) in terms of physical, chemical, biological and microbiological properties. These properties must respect well-defined ranges for ensuring the safety and efficiency of the final formulation. In parallel, a critical study of the key process parameters that could have an influence on the appearance, impurity, and yield of the finished product must be performed. In order to apply properly QbD in process design, it is essential to collect reliable and accurate data for monitoring all possible sources of variability and this is possible thanks to the implementation of Process Analytical Technology (PAT) which ensures the continuous online control of the process itself. FDA has categorized PAT tools into four categories (FDA 2004):

a) Multivariate tools for process design, data acquisition and analysis (Design of Experiments, response surface methodologies, process simulation, and pattern recognition tools)

b) Process analyzers (analytical tools for measuring biological, chemical, and physical attributes at-line, in-line or on-line)

c) Process control tools (complex analytical and statistical strategies developed for defining the attributes of input materials, the ability and reliability of process

analyzers in measuring critical attributes, and the achievement of process end points to ensure consistent quality of the output materials and the final product).

d) Continuous improvement and knowledge management tools (Continuous learning through data collection and analysis over the life cycle of a product for developing approaches and information technology systems that support knowledge acquisition from such databases)

When some or all these tools are appropriately combined, they may be applied to a single-unit operation or to a whole manufacturing process to ensure quality (FDA, 2004). Recent PAT applications involve process monitoring by several advanced process analyzers, such as spectroscopic sensors (e.g., UV-Vis, Near-Infrared and Raman spectroscopic probes) that can acquire a lot of interesting information related to a manufacturing process in a non-destructive way. These techniques generate a large volume of data, requiring the application of chemometrics tools based on multivariate analysis for process monitoring, modeling, and control. Within this scenario, near-infrared (NIR) spectroscopy is considered as a more robust, consistent, and rapid method for the real-time monitoring of complex manufacturing processes, such as granulation [12], drying [13], tablet coating [14] and powder blending [15].

Powder blending is, in fact, one of the key processing steps for ensuring the uniformity and the efficacy of biotechnological products as human and veterinary drugs. In a typical manufacturing process, powder blends of masses anywhere from a few hundred kilograms to tons must be mixed to the point where each unit (typically a few hundred milligrams to a few grams) can be declared to be uniform. The use of NIR sensors allows to collect both chemical and physical information to identify the time at which a mixture is homogeneous and stopping the process. In this way it is possible to perform a real-time control and verification of the whole process, avoiding the need to stop the blender for withdrawing predefined samples for performing off-line chemical analysis. An external mechanical sampling can introduce significant perturbances inside the blender, affecting the quality of a powder formulation and creating unwanted segregation. This is very well explained in a paper published by Esbensen in 2015 [16].

NIR spectra collected along a blending process can be used for developing quantitative and qualitative strategies for the assessment of the blend homogeneity. But before getting into the specific approaches, it is crucial to highlight that the present strategies can only detect when the product is as more homogeneous as possible, according to the blender specifications. In fact, the achievement of the requested mixing level is only related to the design of the blending plant itself that can only be monitored by the PAT.

Quantitative approaches are based on the calculation of a regression model for predicting the amount(s) of an API (Active Pharmaceutical Ingredient) and/or an excipient present in the mixture and considered as a target ingredient. These methods ensure high selectivity to the parameter of interest that can be difficult to assess with qualitative methods. Traditionally, linear methods such as Partial Least Squares regression (PLSr) [17] are applied for maximizing the covariance between the data matrix X of sensor measurements and the matrix Y of parameters to be predicted. However, end-point criteria based on the quantification of a single compound, usually the API, could be not sufficient to assess the uniformity of the whole formulation. Moreover, the external sampling required for the development of the model can dramatically affect the efficiency of the mixing process including in the calculation the significant error caused by the unwanted segregation effect. Moreover, from a more theoretical perspective, the amount of the target ingredient monitored does not change along the process; so, this ingredient is used only as an indicator light for detecting the mixing endpoint. But, adding the numerical value of the predicted quantity can be considered only as mathematical stretch for visualizing, as final outcome of the process, the expected concentration of the key ingredient.

In contrast, qualitative approaches evaluate the evolution of the process as a whole using theNIR fingerprint and its changes, in terms of both shape and intensity over time, as an indicator of the heterogeneity/uniformity of the product itself. The more the NIR signal is stable, the more the product is uniform because is uniform the mixture presented to the sensor at every rotation of the blender. A critical step that needs to be taken into account in the development of such a strategy is the reliability of the outcomes achieved by the NIR sensor due to the possible lack of representativeness of the amount of powder presented to the sensor in respect to the total mass of the

mixture. For considering this key point, is essential to balance the number of acquisitions over time (the faster the process the more frequent the acquisitions) and the number of consecutive acquisitions in specification (the bigger the blender the higher the number of consecutive acquisitions that have to confirm the homogeneity).

Qualitative approaches for assessing the uniformity of a blending process can be based on univariate or multivariate statistics according to degree of specificity required. Moving Block methods (MBM) [18] are hugely applied in blending monitoring due to their simplicity and effectiveness. MBM are based on the calculation of key statistics as standard deviation or mean through a moving window on consecutive blocks of spectra for determining an overall mean or standard deviation plotted against the time. These approaches require a minimum of calibration work and do not suffer from the sampling restrictions imposed when trying to develop quantitative models. More specifically, the Moving Block Standard Deviation (MBSD) allows to summarize the variance contained in a block of spectra down to a single value, which is an indication of the variability of the blend over the time period in which the block of spectra was collected. This is a moving average method, meaning that information of one block is not totally independent of information of the next blocks. In the present study, an alternative application of the Moving Block Standard Deviation MBSD has been proposed for developing a Multivariate Statistical Process Control (MSPC) [6] model in order to monitor the large-scale blending process of a semen extender formulation. This work has carried out thanks to the collaboration of Medi Nova S.a.s., Italian leader company in animal artificial insemination, with the aim to implement directly in the manufacturing plant, a user-friendly platform for real-time process verification and for detecting the endpoint of the mixing. Semen extenders preserve sperm by stabilizing its properties, including sperm morphology, and motility. They must also provide a favorable pH, adenosine triphosphate, anti-cooling and anti-freeze shock, and antioxidant activity to improve semen quality for fertilization [19]. For ensuring all of these properties it is crucial that the active ingredients are present in the right proportion into the formulation and, so, a proper homogeneity needs to be reached.

In the present work, the development of a comprehensive and straightforward strategy for monitoring the blending process of the semen extender is proposed, combining a

NIR miniaturized sensor with a multivariate control chart based on the combination between moving-block standard deviation (MBSD) and principal component analysis (PCA). For a direct interpretation of the multivariate outcomes, a dedicated software was developed, so that process monitoring, and adjustments can be carried out in real-time.

Experimental Plan: Sampling and Spectroscopic Analysis

Semen extender:

Semen extender is a diluent powder formulation which is added to semen to preserve its fertilizing ability. The formulation of the diluter produced by Medi Nova has been designed and developed at the Biopharmanet-Tec Center, a technopole in Parma. The raw materials contained in the diluter are 90% sugars, used as a source of energy for spermatozoa, and the remaining 10% consists of compounds (not declared for industrial secrecy) with high anti-bacterial power and able to adjust the pH in order to preserve the optimal motility and vitality parameters of seminal material.

For a proper development of the process monitoring strategy, a total of 26 independent batches were monitored, in the time range between February 2019 and October 2020. The 26 batches have been divided randomly in two sub-sets: 20 batches have been used for calibration purposes, while the remaining 6 runs have been used to perform validation. In each subset both the years of sampling have been represented in order to avoid possible effects due to time of production and analysis.

Blending details:

After weighting the components, a preliminary sieving has been performed using a circular vibrating screen (Erimaki snc, Italy) for obtaining the optimal particle size. For all the runs, the ingredients have been introduced mechanically through the top following the same filling order, according to a geometric dilution scheme. Mixing has been performed using a powder blender with a fix tank, “four way” model (VIANI snc, Italy) with a capacity of 500 L that was operated at 27 rpm for 20 minutes powder blending Fig 1.

Figure 1: Viani Blender seen from above

Miniaturized NIR device as PAT:

A portable NIR spectrometer, MicroNIR PAT-U (Viavi solutions, Santa Rosa, California USA), which operates in the spectral region 900–1650 nm, was used to acquire the NIR spectra. This instrument, powered (5 V) and controlled via USB port of a computer, originally employs two tungsten light bulbs as the radiation source, a Linear Variable Filter (LVF) as the dispersing element and an uncooled 128-element detector (InGaAs). Spectral resolution is 6.3 nm at 1000 nm. The signals have been obtained by averaging 200 scans and a spectrum has been recorded every 5 seconds. Before starting the measurements, the 0% reference value was taken by leaving the tungsten Lampson with an empty support (known as dark Scan) and an external white reference (Spectralon®) was also scanned to calibrate the device. The instrument has been mounted on the lid of the bin via a sanitary flange with sapphire process window, avoiding direct contact with the product.

Chemometric approach proposed

The data processing strategy can be divided into two steps: the first part was focused on the calculation of standard deviation spectra (SDS) using the Moving Block Standard Deviation method (MBSD) [5]. The final aim of this first step was to remove the contribution of systematic variations that occurred during the mixing, and in particular between different baches, allowing to take into account only the spectral variance related to the process over time. The calculation of a SDS is the crucial step of the proposed strategy because it allows to include the spectral information in the process monitoring, instead of summarizing the variability just with an index (as reported for MBSD). The advantage of considering spectral information, in addition to being more sensitive to changes, lies in the possible interpretation of the signals and so of the reasons why a batch is detected as out of specification.

The second step was the development of a MSPC model based on the application of Principal Component Analysis [6]. For using PCA as a diagnostic tool for process verification, it is necessary to develop the model on data ascribable to the product within specification. So, in the present work, PCA model has been calculated on the calibration batches taking into account only the spectra of the product considered as more homogeneous as possible. These spectra were selected as the last 10% of spectra, time wise, acquired in the last minutes of the process and so, as a good approximation, considered as homogeneous. Thanks to the selection of a suitable number of principal components, statistical boundaries were computed allowing to define a multivariate space describing the mixed product. Calculation of statistical indices based on Hotelling T^2 and Q residuals concluded the strategy with the development of a multivariate control charts to assess whether the mixing is still occurring, or the product can be considered blended. The process can be stopped, and consequently the product can be considered mixed, when 15 consecutive sampling points i.e., spectra acquired at each predefined time interval, fall in the acceptance area of the influence plot. This area is defined by the statistical limits of T^2 and Q residual, chosen according to the predefined confidence level (95% in the present work).

An independent validation set, composed of six independent batches, was used to test the robustness of the strategy. In this case the whole process was taken into account so that the process trajectory in the multivariate space defined by the selected PCs was visualized.

Multivariate data analysis was performed in the Matlab® environment, version 2020a (The MathWorks, Inc., Natick, MA, USA) using in-house functions.

App NIRNova

The final outcome of the present project was the online implementation of a dedicated application developed in MATLAB for the real-time monitoring of the blending process. To make easier the interpretation of the results, the model has been integrated in a user-friendly interface which is shown in Figure 2.

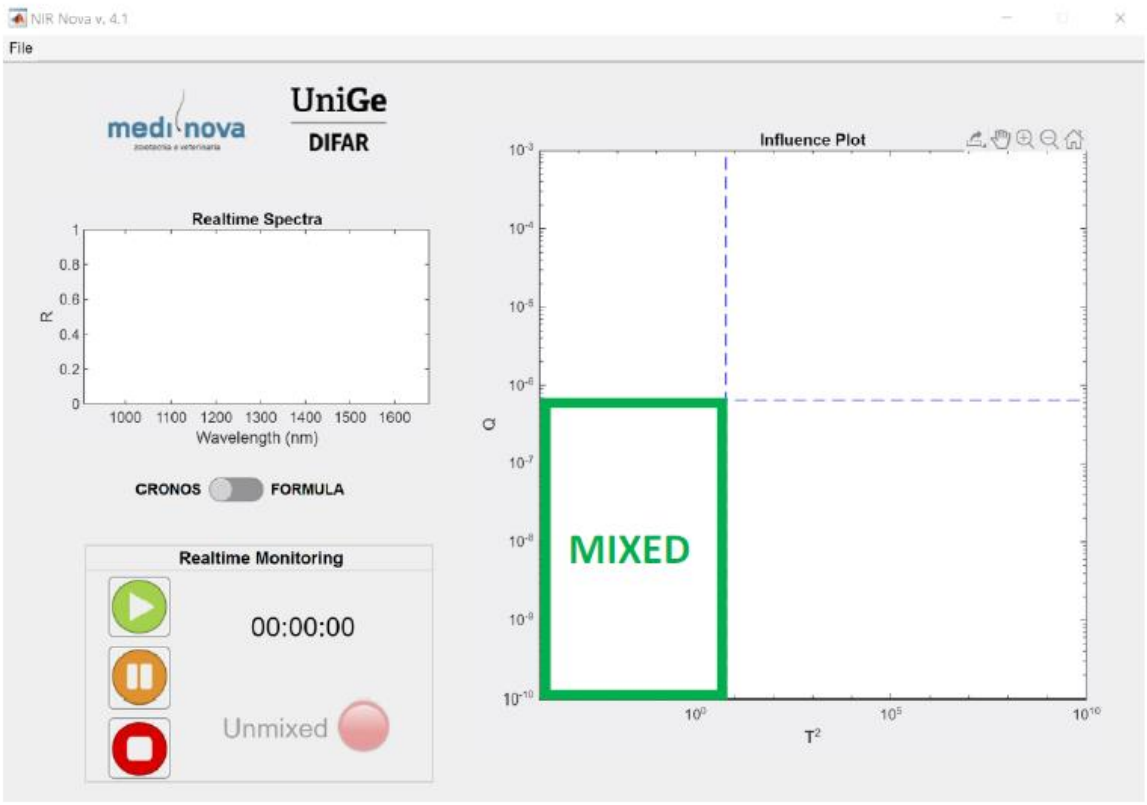

Figure 2: User Interface NIR Nova MATLAB application

Each spectrum measured by the MicroNIR along the blending process is pretreated in according to the preprocessing strategy selected and the MBSD is applied point to point for generating SD spectra using a block size equal to 5. The SD spectra are projected in the space of the PCA global model calculated including all the 26 batches available and applying mean centering as column pretreatment. The projected scores obtained are used to calculate the values of T^2 and Q in order to provide an influence plot as graphical output of the analysis. After the first 5 spectra required for calculating the firs SD profile, a red or green dot appears in the influence plot for each acquisition. The red dots fall outside the limits modeled by the PCA (i.e. with higher values of the T^2 and Q limits) and are related to the moment in which the process is occurring and the product is still “unmixed"; in this situation, in the "Realtime Monitoring" box there is a red light by the wording "unmixed". When the points projected in the influence plot fall inside the limits of the model, the product can be considered mixed, and the blender can be stopped. Trough empirical tests, it was possible to set an endpoint criterion of 15 consecutive spectra that must fall inside the statistical limits in T^2 and Q for considering the mixing completed. At the same time, a green light with the wording "mixed" appears in the "Realtime Monitoring" box. Using the NIR Nova interface, it’s possible to define in a straightforward way the endpoint of the mixing process by performing a real-time process verification of the whole batch.

Moving Block Standard Deviation (MBSD) method

In order to optimize the quality of the raw NIR spectra, six different combinations of row pretreatments have been tested:

1. No row preprocessing
2. Standard Normal Variate (SNV)
3. Savitzky-Golay first derivative (II, 5)
4. SNV + Savitzky-Golay first derivative (II, 5)
5. Savitzky-Golay second derivative (II, 5)
6. SNV + Savitzky-Golay second derivative (II, 5)

The Savitzky and Golay first derivative (Der 1), with a second-degree polynomial order and a window size equal to 5 datapoints, has been used to correct the baseline vertical shifts (offsets) due to temperature variations occurred during the blending process [20]. On the pre-processed data, an alternative application of MBSD has been performed. In this case the standard deviation of the spectra has been calculated over the samples n = 1, ..., N in a selected block. The result is a vector containing the standard deviation calculated at each wavelength according to the formula:

$$\sigma = \sqrt{\frac{\sum_{n=1}^{N}(x_r - \mu)^2}{N-1}}$$

This step generates new profiles that summarize the informative spectral variation without being influenced by unwanted systematic variations between the batches, such as changes in particle size/humidity of the ingredients or variations of the environmental conditions. In order to obtain standard deviation spectra, the MBSD has been applied using the whole spectral region available (125 points) and a block size of 5 which corresponds to the amplitude of the moving window used for the calculation.

Research outcomes

On the calibration set, different row spectral pre-processing techniques have been tested to optimize the quality of the NIR signals (Fig3).

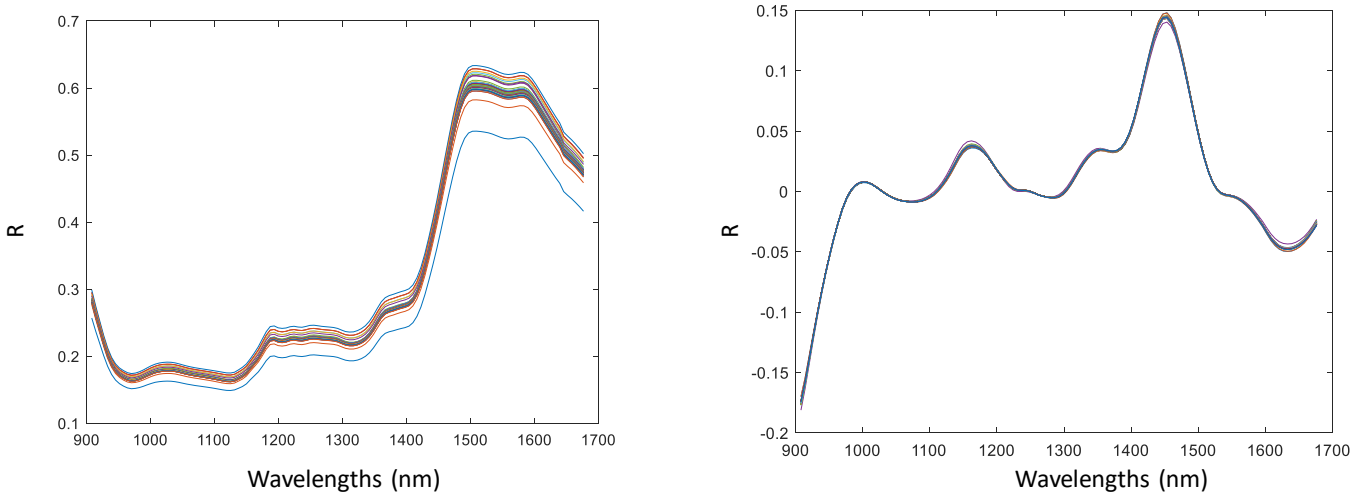

Figure 3: Raw NIR spectra (3a) and pretreated spectra (3b)

On the pre-treated spectra, the Moving Block Standard Deviation (MBSD) has been applied along the sample direction selecting a block size of 5 to obtain SD spectra; in more detail, the standard deviation was calculated at each wavelength for contiguous blocks of 5 spectra. This approach allowed to minimize systematic differences that might be present between batches, due to external factors, focusing on the blending degree. In figure 4 the SD spectra for a calibration batch have been reported

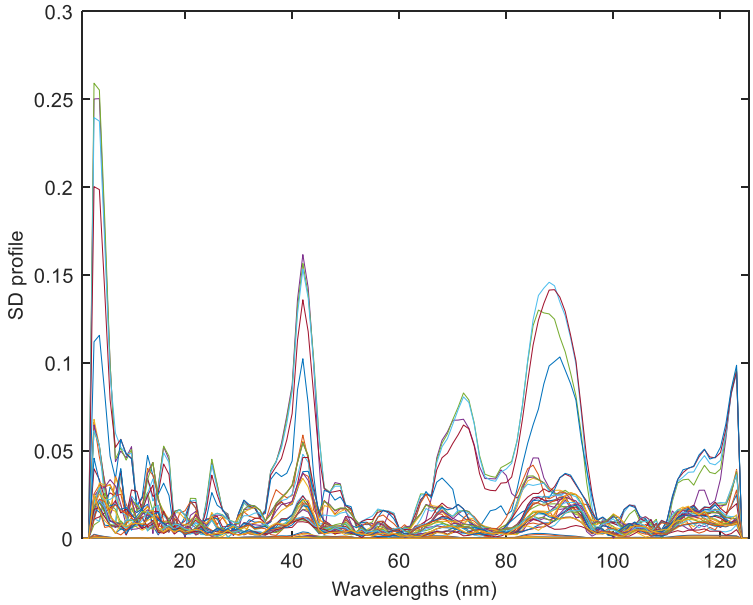

Figure 4: Standard deviation spectra

Finally, the SD spectra related to the calibration set were mean-centered column-wise, before performing a preliminary PCA to study the trajectory of the process and evaluate the consistency among the batches. In Figure 5, the score plot related to the global PCA performed on the calibration set is reported.

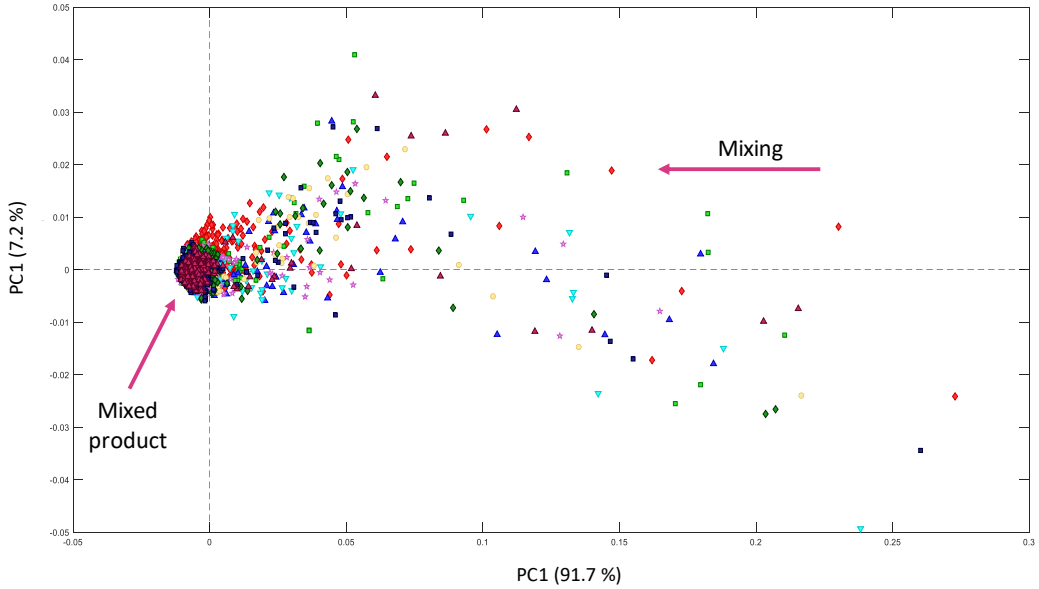

Figure 5: Score plot PC1 vs PC2 related to the global model

The first and the second PCs explained around 98% of the total variance, proving to be the sufficient for describing the system. Along PC1, it is possible to identify, for all the batches included in the calibration set, a clear trajectory related to the mixing process from positive to negative scores values, where the variability among the spectra is minimized. According to these outcomes, it was possible to select a representative calibration set for characterizing the process at the endpoint, including only the last 10 % of the spectra recorded during the blending. A second PCA model was calculated only on the last 10% of the SD spectra, which correspond to the mixed product, in order to define the space related to the moment in which the product can be considered homogenous. The 6 independent validation batches were, then, projected into the space defined by the two lowest-order principal components, accounting for more than 95% of the explained variance. The number of components was decided thanks to a dedicated cross-validation strategy. The influence plot (Hotelling's T^2 vs. Q residuals) and its statistical boundaries at a 95% confidence level were implemented, as a multivariate control chart, in an in-house app developed under MATLAB environment, for the real time monitoring of the behavior of new batches in the orthogonal space defined by PCA. In Figure 6, the influence plots related to the 6 validation batches are shown.

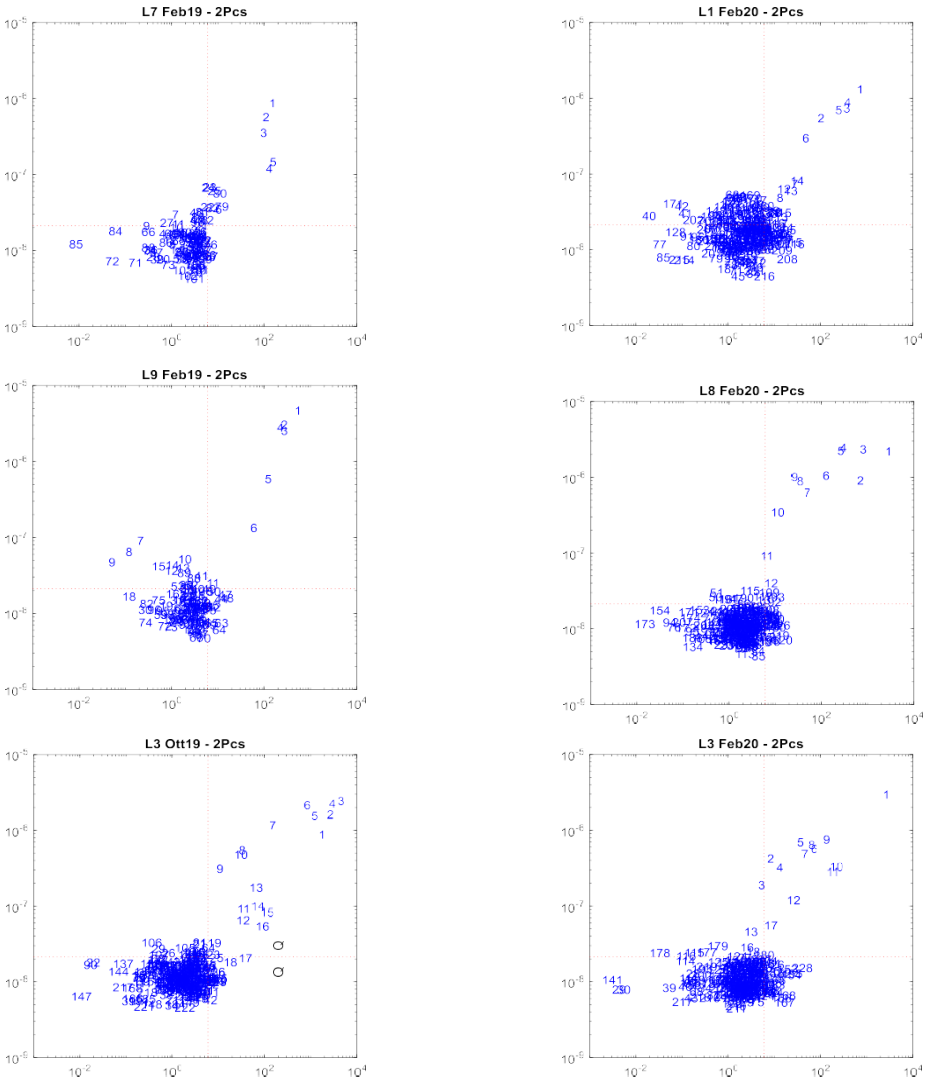

Figure 6: Multivariate control charts on T^2 and Q statistic of six validation batches

The spectra have been numbered in the plots according to the order of acquisition and it is possible to identify a clear trend from the right upper corner to the lower left rectangle, where the limits calculated for both $T^{2,}$ and Q statistics defined the space related to the homogenous product. For all validation batches, regardless of the production year, the product reached this area of the influence plot after about twenty acquisitions, proving the reliability of the approach over the time. The outcomes obtained for the validation set have been confirmed by reference chemical analyses

based on titration methods, allowing to set the endpoint criterion after acceptance of 15 consecutive spectra recognized as compliant by the multivariate control chart.

Conclusion and scientific impacts

The present work, done in collaboration with Medi Nova s.a.s, demonstrated the implementation of a miniaturized NIR sensor, directly online, to be a suitable and powerful tool to perform the real time monitoring of a zootechnical formulation blending and to define the endpoint of the process without requiring invasive sampling, or time-consuming analyses. A deep understanding of the system allowed to account for systematic effects that occur during the mixing process. An ad-hoc 2-step chemometric strategy based on the application of the Moving Block Standard Deviation and the development of a MSPC model allowed to minimize the contribution of these unwanted systematic effects and to build multivariate control charts for a straightforward endpoint detection. Thanks to the present approach, the average time of the process was dramatically reduced from 20 min to 5 min. Moreover, due to the satisfying results obtained, the following chemometric strategy has been implemented in plant thanks to the design of a user-friendly MATLAB interface used for performing the real time endpoint detection of the blending process.

3.2. POWDER BLENDING MONITORING BY MINIATURIZED NIR SENSOR: A CRITICAL COMPARISON OF MULTIVARIATE QUALITATIVE APPROACHES FOR UNIFORMITY ASSETMENT

Scientific background and aim of the work

Powder blending is a unit operation that represents one of the key processing steps for ensuring uniformity of a final formulation. In a food or pharmaceutical manufacturing process, powder blends of masses anywhere from a few hundred kilograms to tons must be mixed to the point where each unit (typically a few hundred milligrams to a few grams) can be declared to be uniform.

To reach this goal, it is crucial to optimize the particle size, texture and cohesiveness of both major and minor components of a mixture, the parameters of the process as bin filling, temperature, rotation speed and mixing time in order to achieve the desired endpoint of uniformity avoiding unwanted effects as powder segregation [21]. The traditional approach for blending monitoring is carried out discontinuously, i.e., stopping the blender after a certain time, usually defined from experience, and verifying the process by withdrawing 10 samples in a predefined position (specified by current ICH). The samples extracted from the bin are then analyzed by reference methods as UV\VIS or HPLC off-line in the quality laboratory. However, interruption of the process and withdrawing of samples for evaluating blend uniformity may cause big perturbances to the system, leading to powder segregation and, thus, affecting the efficiency of the mixing process [16].

In this scenario, there is a great deal of interest in applying emerging process analytical technologies (PATs) for improving understanding of blending processes with the aim to perform a real-time monitoring. Near Infrared Spectroscopy (NIRS) proved to be a very accurate and efficient tool for acquiring in a no-destructive way every step of every batch during the production phase [3]. From a practical point of view, an NIR sensor is usually mounted to the surface of a blender, typically on the lid and collects a single spectrum of an essentially static sample during each rotation of the blender.

It possible to identify an effective scanning zone (Field-of-View, FoV) associated with the instrument, which is a function of [22]:

- Blender rotation speed.
- Instrument integration time.
- Number of scans to coadd to generate a single spectrum.
- The signal-to-noise (S/N) of the instrument is a function of the number of coadded scans, therefore it is important to maximize this number based on the rotation speed of the blender.

All these parameters have been considered for defining the experimental plan and the acquisition setting of a real industrial case-study. In this work, thanks to a long-term collaboration with Viavi Solutions, it was possible to implement a MicroNIR PAT solution on a large-scale blender for the monitoring of the mixing process of the multicomponent solid fraction of a commercial energy drink. On the spectroscopic information collected along seven independent batches, three unsupervised qualitative chemometric methods have been tested and compared for assessing the homogeneity of the mixture and evaluating the conformity of the process according to HPLC results. The data processing was carried out in collaboration with the University of Barcelona during my internship abroad.

Experimental plan, sampling, and spectroscopic analysis

The formulation investigated included six different ingredients: caffeine, sugars, nicotinamide, taurine and vitamin B6 (NDA does not allow to report all the formulation details). In according to the company guidelines, nicotinamide is considered the target ingredient to be monitored in the blend, due to the fact it was one of the minor components constituting 2.2% of the product.

At the end of each run, the reference analysis based on the HPLC method was performed in the quality laboratory of the company for quantifying the content of

nicotinamide. The samples were taken during blend discharge. For each top, middle and bottom sections of the blender, four samples were taken (12 samples in total) and the HPLC results certified that batches 1-2-3-4-5-7 were compliant, with the only exception of the sixth run which could not be considered compliant according to the final amount of nicotinamide.

The experiments have been conducted in two consecutive days, varying the fill level and the loading mode in order to evaluate possible effects of these parameters. In Table 1, the experimental plan has been reported.

Table 4: Experimental plan

Batch	Compliance	Day	Fill Level	Loading
Batch 1	YES	30-07-2021	65%	Manually
Batch 2	YES	30-07-2021	65%	Manually
Batch 3	YES	30-07-2021	65%	Automatically
Batch 4	YES	1/8/2021	45%	Automatically
Batch 5	YES	1/8/2021	45%	Automatically
Batch 6	NO	1/8/2021	45%	Manually
Batch 7	YES	1/8/2021	45%	Manually

The tumble blender (Cyclops Maxi, IMA, U.S.A) used for the analysis had a capacity of 2000 L and the mixing process was performed for all the batches with a rotation speed of 8 rpm for 15 minutes, according to the protocol of the company.

The spectral information recorded in real-time for each batch has been collected by a miniaturized NIR sensor (MicroNIR© PAT-W, Viavi Solutions, Santa Rosa, U.S.A) implemented on the lid of the blender trough a triclamp flange. During the mixing, NIR spectra have been collected through a sapphire window, thus avoiding any contact with the sample. The signals have been recorded in diffuse reflection mode with an

integration time of 12.3 milliseconds, this means that a single NIR spectrum can be potentially measured every 1.23 seconds, considering the whole spectral range available (900-1700 nm). In order to increase the quality of the signals, a scan count of 100 has been applied for calculating each signal as the averaging of one hundred independent scans. For this application, a NIR spectrum has been recorded every blender rotation on the gravity sensor of the wireless device ensuring the measurement of a representative sample.

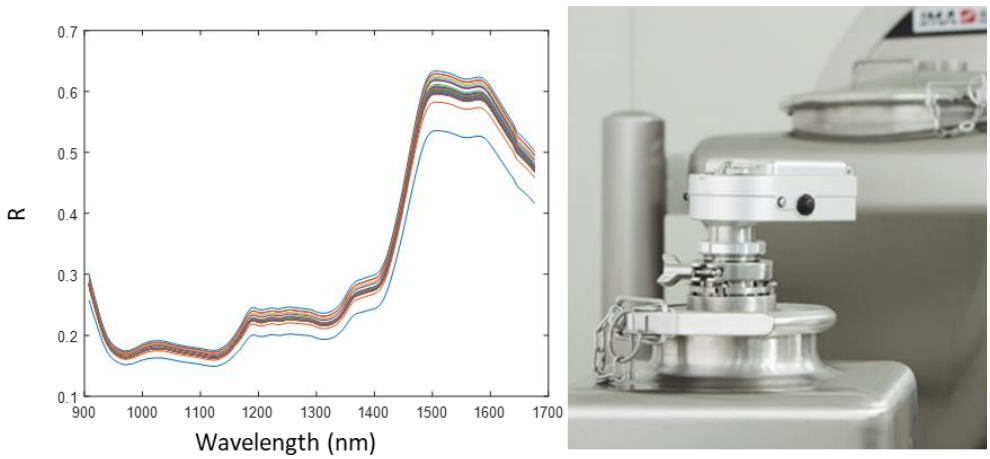

Figure 8: Raw spectra (left) and MicroNIR PAT-W implemented on the blender (right)

Chemometric Approach proposed

The NIR spectra have been collected by using the MicroNIR Pro Software and the data processing was performed in Matlab environment (The MathWorks, Inc., Natick, MA, USA, Version 2020b) using both the PLSToolbox package (Eigenvector Research, Inc. Manson, Washington) and in-house functions. In order to assess the uniformity of the formulation, three qualitative methods have been tested:

- Moving Block F-Test [23]
- Multivariate Statistical Process Control (MSPC) based on Principal Component Analysis (PCA) [6]
- Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) [7]

Each method has been applied separately on the preprocessed NIR spectra acquired along the seven independent batches considered in the present case-study. The results obtained, in terms of endpoint detection and evaluation of the conformity of each run, have been discussed to provide a critical comparison about the selected chemometric strategies.

Moving Block F-test

Moving block methods are used for end-point detection where the purpose is to find when the process has stabilized (i.e., is stationary). The Moving F-Test is based on Fishers F-Test where a 95% confidence limit can be used to statistically compare the variances calculated within contiguous blocks of spectra that are independent from each other; the process is considered stabilized when the F-test does not show a significant difference in variances of contiguous blocks. The variance for the two blocks of spectra is calculated by performing an F-test using N-1 degrees of freedom both at the numerator and at the denominator, according to the formula:

$$F_{B_x/B_{x-1},0.05} = \frac{s^2_{B_x}}{s^2_{B_{x-1}}}$$

The F-plot obtained against the time can be used for end-point detection as the moment in which the variance decreases, and the curve becomes smooth [23].

- *Multivariate Statistical Process Control (MSPC) based on Principal Component Analysis (PCA)*

Multivariate statistical process control (MSPC) models aim at providing statistical boundaries that allow to build multivariate control charts to assess whether a process is on- or off-specification based on a set of continuous measurements. MSPC models based on PCA can be applied for different goals, such as batch endpoint detection or

checking process evolution. In this work, a PCA-MSPC based model has been developed for the endpoint detection of the blending process. Other details about this approach are reported in the previous paragraph (3.1).

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

Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) provides the concentration profiles and pure spectral fingerprints for all compounds involved in a mixture by using only the spectra acquired during the process evolution. By the calculation of a bilinear model represented by the following equation, MCR-ALS provides physically and chemically meaningful concentration and spectral profiles of the pure components of the system:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E}$$

where $\mathbf{X}$ is the original data matrix with spectroscopic process observations; $\mathbf{S}^{\mathbf{T}}$ contains the pure spectral signatures of the components needed to describe the process and $\mathbf{C}$ the related concentration profiles. $\mathbf{E}$ is the matrix with the residual part not explained by the model related to the experimental error.

MCR-ALS obtains $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$ matrices using an alternating iterative optimization method. First, an initial estimate of $\mathbf{C}$ or $\mathbf{S}^{\mathbf{T}}$ should be used to start the iterative procedure. Then, in each iterative cycle, the $\mathbf{C}$ and $\mathbf{S}^{\mathbf{T}}$ matrices are calculated according to preselected constraints, applied to reduce the rotational ambiguity of the final solutions and to give physicochemical meaning to the profiles retrieved (Tauler et al., 1995). The optimization continues until a predefined convergency criterion is met.

In this work, the concentration profiles calculated by MCR-ALS have been used to identify the endpoint of the blending process. This approach allowed us to evaluate the evolution of the mixing process ingredient by ingredient, with a focus on nicotinamide, which was considered as the target compound of the mixture. According to the nominal concentration of nicotinamide (2.2%) a scale factor for converting the arbitrary unit of the concentration profile to the real concentration unit (%) has been calculated, providing a semi-quantitative solution.

Research outcomes

In order to optimize the quality of information provided by NIR spectra, a preprocessing optimization has been performed. Three different combinations of pretreatments have been tested: Standard Normal Variate (SNV) transform, Savitzky–Golay first derivative (5 datapoint window, second polynomial order) and normalization (using 2-norm, i.e., the Euclidean norm), Savitzky–Golay second derivative (5 datapoint window, second polynomial order). The second strategy, which included SavGol first derivative and normalization, proved to be the best one according to interpretability of the concentration profiles obtained by the following MCR-ALS analysis.

On the pretreated signals, the Moving Block F-Test was applied on a calibration set which included the first five batches resulting compliant according to the HPCL reference analysis. The model obtained with a block size of 10 has been validated using the data related to the other two runs, the sixth (non-compliant) and the seventh (compliant).

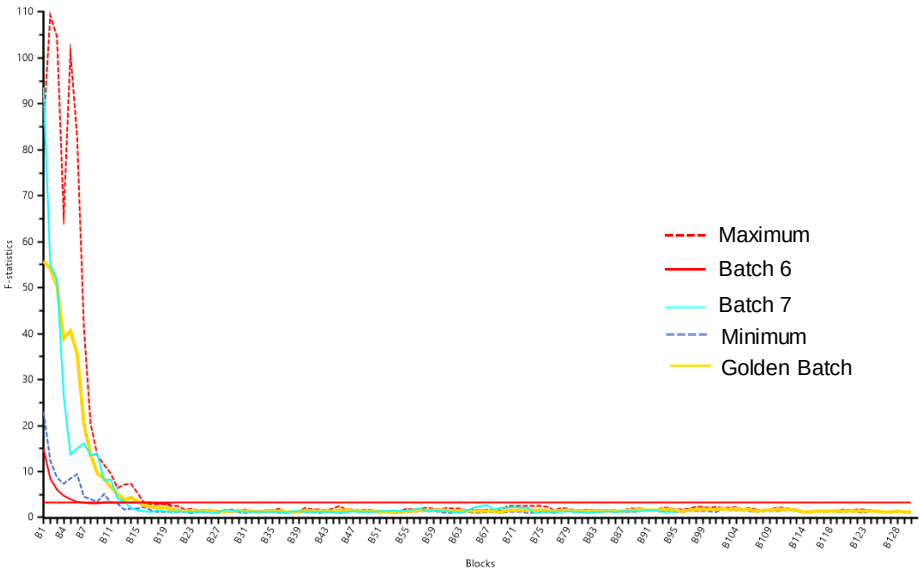

Figure 3: Results of validation of the Moving Block F-Test

Figure 3 shows the sixth (red profile) and seventh (ligh blue) batches projected in the space of the MB model calculated on the calibration set. Thanks to the golden batch visualization it is possible to define the limits of the experimental domain: the red dashed profile indicates the upper limit and the blue one corresponds to the lower limit. The golden batch profile is obtained by averaging the Moving Block profiles calculated on the compliant batches included in the calibration set; the golden batch can be considered as the optimal standard. Moreover, the algorithm automatically determines the limit based on the Fisher statistics, which corresponds to the endpoint of the blending process. In accordance with HPLC results, batch 6 falls outside the limits of the domain at the beginning of the mixing process and it reaches an earlier endpoint, in respect to the golden batch, around block 11. This run shows a different evolution in respect to the compliant batches, but the differences can be appreciated only partially at the first rotations of the mixing process. Batch 7, considered compliant for HPLC analysis, is within the limits of the experimental domain from the beginning to the end, achieving the endpoint around block 15, according to the evolution of the golden batch profile. These outcomes demonstrated the suitability of the Moving Block F-Test in the monitoring of the spectral variability in order to detect the endpoint of blending process. However, this approach does not allow to perform a real diagnostic for assessing the conformity of a further run and does not provide any specific information about the target ingredient (Nicotinamide) proving to be less sensitive to low dosage formulations.

Following the previous strategy, the MSPC model based on the PCA is developed on a calibration set which included only the first five compliant batches. The PCA model is calculated selecting only the last 10% of the spectra acquired along the process, which correspond to the endpoint. In this way it is possible to define the space of the model related to the blended and homogenous product. This approach allows to build multivariate control charts based on Hotelling's T^2 and Q residuals, which provide statistical limits for the identification of the endpoint. In order to summarize the overall variability of the process in a single control chart, in Figure 4a T^2–MSPC chart is shown. The seventh compliant batch (in green) shows higher T^2 values at the beginning of the process that decrease according to the evolution of the mixing until reaching the endpoint at spectrum 180, after around 15 minutes.

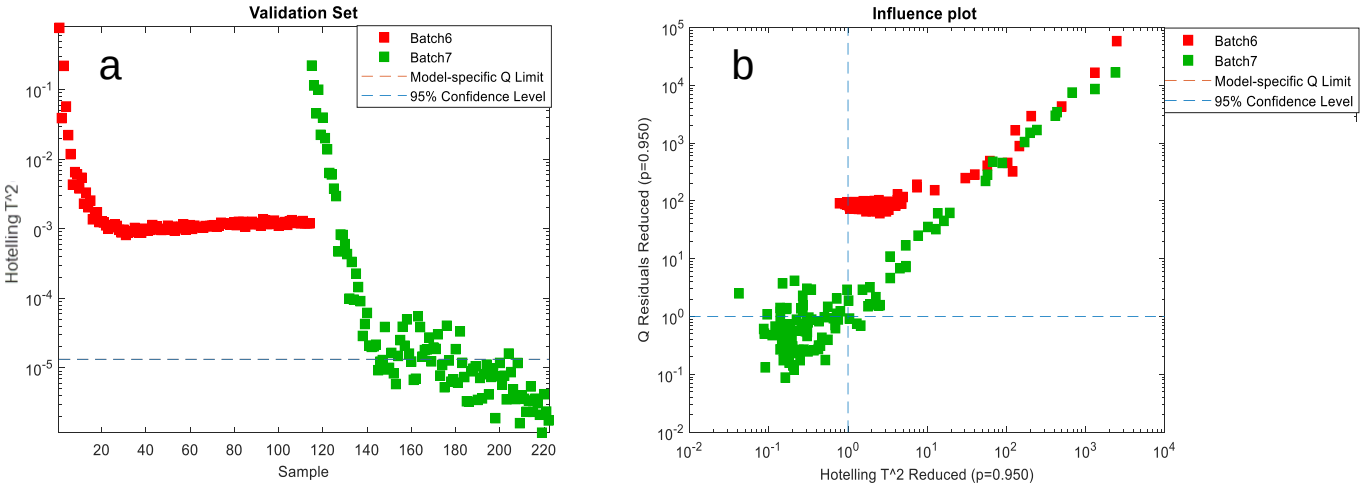

Figure 4: (4a) shows the values of T^2 along the mixing for the validation set. (4b) shows the influence plot (Q-residuals vs. T^2) related the same validation batches

However, the non-compliant batch, represented in red, stands at higher T^2 values throughout the duration of the blending without reaching the endpoint. Figure 4b shows the influence plot built on both Hotelling's T^2 and Q residuals, which confirms the previous outcomes. At the end point, batch 7 is inside the limits of the model proving to follow the same behavior of the compliant calibration batches. However, batch 6 seems to follow a similar evolution achieving a sort of stabilization but it falls outside the limits of the model. In light of these results, the MSPC approach can be considered a more powerful diagnostic tool in respect to MB F-Test method, allowing at the same time to define objective statistic limit for the endpoint detection.

The last method applied on this data set is based on the application of the MCR-ALS method. The model has been calculated on preprocessed data after application of non-negativity constraints in concentration profiles. For each batch, it was possible to obtain the concentration profiles of the six ingredients inside the formulation in order to study the evolution of each single ingredient along the mixing process (Figure 5 a). Moreover, to evaluate the consistency of these results in respect to the HPLC analysis, the concentration profiles of nicotinamide, obtained in an arbitrary unit, have been

scaled to the real concentration unit (%). In this way, according to the acceptance limits of the company (RDS ≤ 5%), it was possible to evaluate the evolution of the only nicotinamide, obtaining a semi-quantitative answer for the target compound. (Figure 5b).

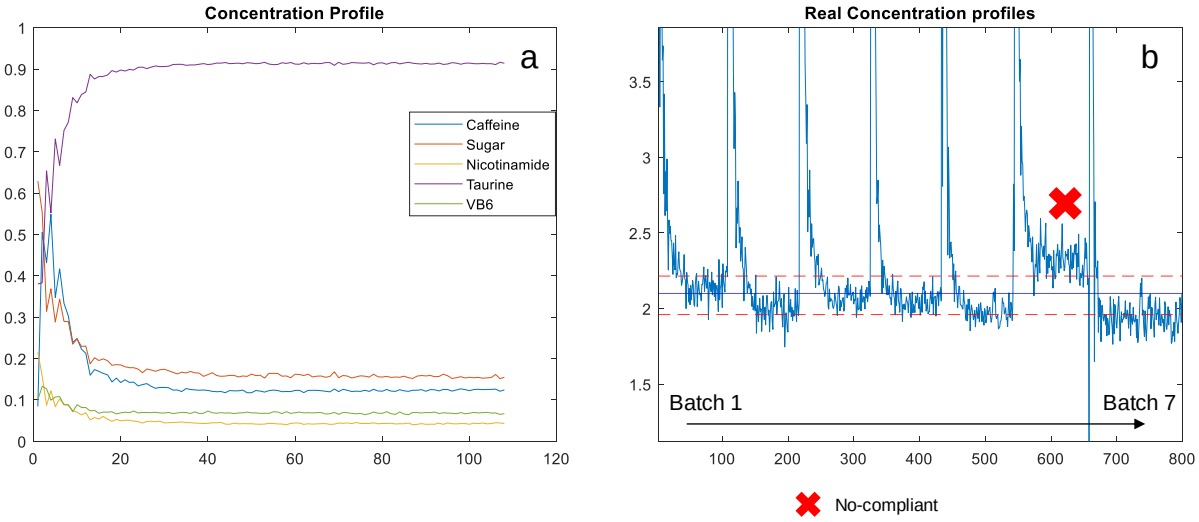

Figure 5: (5a) shows the concentration profiles obtained by MCR analysis for a compliant batch. (5b) shows the Nicotinamide profiles scaled in real unit

In respect to the ideal amount of nicotinamide, equal to 2.2%, the company considers acceptable a relative standard deviation ≤ 5% and these limits have been plotted in Figure 5 b. As it is possible to see in the graph, the concentration profile of nicotinamide related to batch 6 does not reach the desired amount at the endpoint, confirming the HPLC outcomes. MCR proved to be a very powerful technique allowing to monitor low dosage formulations and providing specific information about single ingredients of the formulation. Moreover, some variation of this algorithm, such as MCR-ALS (Multivariate Curve Resolution Alternative Least Square), can provide a quantification of one or more target compounds.

Conclusion and scientific impacts

In this work, a real industrial case-study has been used to evaluate the pros and cons of three qualitative methods for the real-time monitoring of a powder blending process. MB F-test allowed to identify the endpoint of the process through the global evaluation of the spectral variability over time, providing an objective limit based on Fisher statistics and ensuring a straightforward interpretation of the outcomes. This method did not give information about specific ingredients inside the mixture and was not able to perform a real diagnostic for assessing the compliance of production batches. MSPC model based on the application of PCA proved to be a stronger diagnostic tool, which allowed to build multivariate control charts useful for obtaining a comprehensive understanding of the mixing process, for identifying the endpoint according to statistics parameters (Hotelling's T^2 and Q residuals) and for evaluating the compliance of further batches in a more reliable way. However, also this approach proved to be less sensitive to low dosage formulations due to the fact that it is not possible to extract the information related to specific ingredients. MCR-ALS is the only approach that allowed to characterize the evolution of each single component of the mixture through the calculation of concentration profiles. Although this technique has the higher sensitive to low dosage formulations, it does not automatically calculate any statistical limit for detecting the endpoint of the blending process. Moreover, the implementation of the model can be complicated because of the need to optimize many parameters (constraints) for obtaining qualitative or quantitative answers. In general, the results obtained for this specific case-study, demonstrated as each technique can provide a complementary information useful for understanding and modelling different aspects of the blending process. A multistep chemometric strategy based on the consecutive applications of MB F-test, MSPC and MCR-ALS can be considered as a very promising strategy for the real-time and online monitoring of a process.

[1] Siesler, H. W., Kawata, S., Heise, H. M., & Ozaki, Y. (Eds.). (2008). Near-infrared spectroscopy: principles, instruments, applications. John Wiley & Sons.

[2] Beć, K. B., Grabska, J., & Huck, C. W. (2021). Principles and applications of miniaturized near-infrared (NIR) spectrometers. Chemistry–A European Journal, 27(5), 1514-1532.

[3] Woodcock, J. (2004). The concept of pharmaceutical quality. American Pharmaceutical Review, 7(6), 10-15.

[4] De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J. P., & Vervaet, C. (2011). Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. International journal of pharmaceutics, 417(1-2), 32-47.

[5] Momose, W., Imai, K., Yokota, S., Yonemochi, E., & Terada, K. (2011). Process analytical technology applied for end-point detection of pharmaceutical blending by combining two calibration-free methods: simultaneously monitoring specific near-infrared peak intensity and moving block standard deviation. Powder technology, 210(2), 122-131.

[6] Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. International Journal of adaptive control and signal processing, 19(4), 213-246.

[7] Jaumot, J., de Juan, A., & Tauler, R. (2015). MCR-ALS GUI 2.0: New features and applications. Chemometrics and Intelligent Laboratory Systems, 140, 1-12.

[8] Food and Drug Administration. (2004). Guidance for industry, PAT-A framework for innovative pharmaceutical development, manufacturing and quality assurance. http://www. fda. gov/cder/guidance/published. html.

[9] Food and Drug Administration. (2006). Guidance for Industry, Q8 Pharmaceutical Development. Food and Drug Administration: Silver Spring, MD, USA.

[10] US Food and Drug Administration. (2006). Guidance for industry: Q9 quality risk management. Bethesda, MD.

[11] Corredor, C. C., Lozano, R., Bu, X., McCann, R., Dougherty, J., Stevens, T., ... & Shah, P. (2015). Analytical method quality by design for an on-line near-infrared method to monitor blend potency and uniformity. Journal of Pharmaceutical Innovation, 10(1), 47-55.

[12] Alcalà, M., Blanco, M., Bautista, M., & González, J. M. (2010). On-line monitoring of a granulation process by NIR spectroscopy. Journal of pharmaceutical sciences, 99(1), 336-345.

[13] Morris, K. R., Stowell, J. G., Byrn, S. R., Placette, A. W., Davis, T. D., & Peck, G. E. (2000). Accelerated fluid bed drying using NIR monitoring and phenomenological modeling. Drug development and industrial pharmacy, 26(9), 985-988.

[14] Tabasi, S. H., Fahmy, R., Bensley, D., O'Brien, C., & Hoag, S. W. (2008). Quality by design, part II: application of NIR spectroscopy to monitor the coating process for a pharmaceutical sustained release product. Journal of pharmaceutical sciences, 97(9), 4052-4066.

[15] Cuesta Sanchez, F., Toft, J., Van den Bogaert, B., Massart, D. L., Dive, S. S., & Hailey, P. (1995). Monitoring powder blending by NIR spectroscopy. Fresenius' journal of analytical chemistry, 352(7), 771-778.

[16] Esbensen, K. H., Román-Ospino, A. D., Sanchez, A., & Romañach, R. J. (2016). Adequacy and verifiability of pharmaceutical mixtures and dose units by variographic analysis (Theory of Sampling)–A call for a regulatory paradigm shift. International journal of pharmaceutics, 499(1-2), 156-174.

[17] Berntsson, O., Danielsson, L. G., Lagerholm, B., & Folestad, S. (2002). Quantitative in-line monitoring of powder blending by near infrared reflection spectroscopy. Powder Technology, 123(2-3), 185-193.

[18] Blanco, M., Bañó, R. G., & Bertran, E. (2002). Monitoring powder blending in pharmaceutical processes by use of near infrared spectroscopy. Talanta, 56(1), 203-212.

[19] Bustani, G. S., & Baiee, F. H. (2021). Semen extenders: An evaluative overview of preservative mechanisms of semen and semen extenders. Veterinary World, 14(5), 1220.

[20] Oliveri, P., Malegori, C., Simonetti, R., & Casale, M. (2019). The impact of signal pre-processing on the final interpretation of analytical outcomes–A tutorial. Analytica chimica acta, 1058, 9-17.

[21] He, X., Han, X., Ladyzhynsky, N., & Deanne, R. (2013). Assessing powder segregation potential by near infrared (NIR) spectroscopy and correlating segregation tendency to tabletting performance. Powder technology, 236, 85-99.

[22] Vanarase, A. U., Alcalà, M., Rozo, J. I. J., Muzzio, F. J., & Romañach, R. J. (2010). Real-time monitoring of drug concentration in a continuous powder mixing process using NIR spectroscopy. Chemical Engineering Science, 65(21), 5728-5733.

[23] Fonteyne, M., Vercruysse, J., De Leersnyder, F., Besseling, R., Gerich, A., Oostra, W., ... & De Beer, T. (2016). Blend uniformity evaluation during continuous mixing in a twin screw granulator by in-line NIR using a moving F-test. Analytica chimica acta, 935, 213-223.

CHAPTER 4: NIR HYPERSPECTRAL IMAGING

Hyperspectral imaging (HSI) in the near infrared (NIR) spectral region is an innovative analytical tool which allows to acquire the spectral information for a large number of contiguous spatial portions (pixels) related to the surface of the sample analyzed. The real advantage, in respect to the traditional NIR spectroscopy, is the possibility to obtain a spatial representation of the distribution of chemical components (chemical imaging) for qualitative and quantitative purposes [1]. A hyperspectral imaging system produces a two-dimensional spatial array of vectors, which represent the spectrum at each pixel location; the resulting three-dimensional matrix (the so-called hypercube) includes two spatial and one spectral dimension. For reducing the dimensionality of the data and extracting both chemical and spatial information, the application of chemometric techniques is needed. Before starting data modeling, it is crucial to perform some preliminary steps for properly handling hyperspectral images [2]:

1) Removal of noisy spectral data: in correspondence of the extremes of the spectral range, it is possible to find noisy parts that can affect the effectiveness of data analysis.

2) Removal of image background: in most of the cases, hyperspectral images include a lot of not interesting information related to the background that is usually removed choosing the best strategy according to its specific features.

3) Removal of “dead pixels” and spikes: isolated bad pixels placed randomly can generate black pixels, called “dead pixels”, and false intensity peaks known as “spikes”, which should be removed before applying any chemometric strategy.

4) Spectral preprocessing: mathematical pretreatments are applied for optimizing the quality of the spectral information and for avoiding global intensity effects due to light scattering.

Once clean images are obtained, it is possible to apply unsupervised techniques as PCA for summarizing and exploring the information contained in hypercubes. According to the specific aim of the study, PCA can be applied on whole images, selected objects or pixels. In the first case, the analysis is performed considering the whole image as a sample and comparing images to each other's for pattern recognition purposes. The object-based approach considers specific objects of interest inside the image, while the pixel-based approach is focused on individual pixels for emphasizing differences among pixels [3].

A pixel-based level approach has been used in the study described in this last chapter for mapping the evolution of dehydration, proteolysis, and lipolysis during the ripening of a semi-hard cheese.

Chapter 4

4.1 AN IN-DEPTH STUDY OF CHEESE RIPENING BY MEANS OF NIR HYPERSPECTRAL IMAGING: SPATIAL MAPPING OF DEHYDRATION, PROTEOLYSIS AND LIPOLYSIS

Scientific background and aim of the work

In the last 20 years, with the development of technology and chemometric tools, near-infrared spectroscopy (NIRS) established itself as an accurate non-destructive technique for compositional analysis and quality evaluation for a wide range of dairy products including: liquid or dried milk, cream and traditional and processed cheese [4]. Among all dairy products, cheese represents one of the most complex milk derivatives for the high number of factors contributing to its chemical composition and technological characteristics. Despite cheese complexity, NIR spectroscopy has proven to be a suitable analytical method to verify the authenticity of certified products [5] to evaluate cheese chemical composition to follow the shelf-life [6], to determine sensorial parameters [7] and to characterise cheese ripening [8,9]. Indeed, traditional bench-top instrumentation, equipped with point-based scan systems, does not provide information regarding the spatial evolution of phenomena that occur in the product – a crucial aspect that has to be taken into account for inhomogeneous matrices such as many dairy products, cheese above all.

Recent advances are focused on the development of efficient analytical approaches based on the employment of NIR hyperspectral imaging (NIR-HSI) for food analysis, in order to merge the spectral and spatial information and to evaluate, simultaneously, both chemical composition and physical features of the sample [10]. In a hyperspectral image, in fact, each pixel corresponds to a spectrum in the whole NIR range; in this way it is possible to combine the information about chemical composition and the one related to analyte spatial distribution [11]. The advantages of NIR-HSI, compared to conventional NIR spectroscopy, were demonstrated for the prediction of macronutrient content (proteins, fat and carbohydrates) in different cheese varieties [12]. Furthermore, in recent studies, the employment of this technique, combined with chemometric strategies, provided a satisfactory estimation of the rind percentage in

hard Italian ground-cheese [13] and for the prediction of grade ripening of long-ripening cheeses [14].

In the present study, NIR-HSI was tested as an efficient non-destructive tool to follow the biochemical evolutions that occur in the cheese wheel during the ripening/maturation phase. For the first time, the contribution of each of the three most important processes that are involved – lipolysis, proteolysis and surface dehydration – was deconvolved. Thanks to this deconvolution strategy, it was possible to map and study over time the changes ascribable separately to proteins, lipids and moisture. Ripening, in fact, represents a crucial step in cheese making, which involves a balanced series of consecutive microbiological and biochemical events that lead to the characteristic taste, aroma and texture of each cheese variety. In more detail, lipolysis is responsible for the catabolism of triacylglycerol (TAG), catalysed by the action of indigenous, endogenous and/or exogenous lipases. This process leads to the formation of volatile aromatic compounds, contributing significantly to the flavour of many cheese varieties [15]. Proteolysis is the most complex biochemical event that occurs during ripening, with a major impact on flavour and texture; it can be divided into three phases: proteolysis in milk before cheese manufacture, the enzymatically induced coagulation of milk, and proteolysis during cheese ripening [16].

To achieve the objective of an in-depth understanding of cheese ripening, the present study is organised in two consecutive steps.

The first one, aimed at a fundamental understanding of the NIR spectral bands ascribable to each biochemical change, was carried out acquiring HSI data from different cheese varieties, in which proteolytic, lipolytic and dehydration changes occur at a different relative intensity.

Secondly, the maturation of a typical Italian cheese, from the Liguria region, commercially named Formaggetta and provided by Caseificio Val D'Aveto (Rezzoaglio, GE, Italy), was followed. Formaggetta is a semi-hard cheese characterised by a short maturation time; therefore, visualisation of the spatial changes that occur during ripening becomes crucial for a reliable understanding of the phenomenon. Nowadays, indeed, it is of great interest, both for cheese manufacturing and associations involved in the protection of the certified products (consortia) to

dispose of a fast, non-destructive and sensitive method to monitor dynamic transformations during maturation, in relation to the quality of cheese.

In order to manage the high amount of information embodied in HSI-NIR data, a chemometric approach is required. The present study proposes the application of principal component analysis (PCA) for both understanding the predominant NIR spectral bands involved in cheese ripening and creating distinct chemical maps of the main three biochemical events [17].

Experimental Plan: Sampling and Spectroscopic analysis

The experimental plan can be summarized in the following two steps:

STEP A – biochemical process understanding: this step, aimed at identifying the NIR spectral bands that characterize lipolysis, proteolysis and dehydration, involved analysis by means of HSI-NIR of different cheese types. In more detail, a sample set including 12 commercial varieties of cheese, was considered:

• Asiago, Casera, Parmigiano Reggiano 20 months-aged and 30 months-aged were selected for being characterised by proteolytic maturation (Atlante sensoriale dei prodotti alimentari, 2012), at different intensities.

• Avetino, Camoscio d'Oro, President, Gorgonzola and Roquefort were collected in order to investigate lipolytic reactions, which are predominant during their ripening (Atlante sensoriale dei prodotti alimentari, 2012).

• Morbidezza, Formaggetta and Bel Paese, were included in the sample set for their intermediate biochemical behaviour (Atlante sensoriale dei prodotti alimentari, 2012).

Formaggetta and Morbidezza samples were provided by Caseificio Val d'Aveto (Rezzoaglio, GE, Italy), while the other commercial samples were purchased at a local market. All the samples were sliced by the cheese seller (2 cm thick), few hours before analysis. Samples were let equilibrating at room temperature for one hour, prior to image acquisitions, to avoid temperature interferences in the NIR signals. All samples were measured without any chemical pre-treatment/preparation.

STEP B – biochemical mapping over time: in the second phase, the ripening of Formaggetta cheese was studied from a chemical point of view, mapping lipolysis, proteolysis and dehydration during the last 10 days before market release. For this step, a total of twenty samples (completely independent from the ones used in STEP A) were produced in a single batch and delivered by Caseificio Val d'Aveto, two for each of the ten sampling dates, to the Department of Pharmacy (DIFAR) of the University of Genova for the analyses. To better represent the spatial evolution of cheese ripening, every day one Formaggetta wheel was sectioned longitudinally, while the other one transversally, with a thickness of 2 cm in both of the cases. In this way, a total of twenty hyperspectral images of independent samples were acquired over the whole ripening process. Hyperspectral images were acquired using a line scanning system (Specim Ltd, Finland) composed by a SWIR3 hyperspectral camera, working in the 1000-2500 nm spectral range at 5.6 nm resolution. The camera is equipped with three halogen lamps (35 W, 430 lm, 2900 K, each) at a 45° incident angle as the illumination source, and a horizontal line scanner (Lab Scanner, 40 × 20 cm moving stage). The system was controlled by the Lumo Scanner v. 2.6 software (Specim Ltd, Finland). Prior to each measurement, dark (closed shutter) and white (99% reflectance Spectralon® rod) images were automatically recorded and stored.

A total of 32 images were acquired: one image for each of the 12 commercial cheeses and two images/day, for a total of 10 days representing the Formaggetta ripening process.

Chemometric approach proposed

The first step of the hyperspectral images processing was an internal calibration to normalise the reflection intensity values recorded (I) in the raw images according to the white (W) and the dark (D) intensities, for each pixel p and at each wavelength λ, independently, obtaining reflectance values (R) according to the following equation:

$$R_{p,\lambda} = \frac{I_{p,\lambda} - D_{p,\lambda}}{W_{p,\lambda} - D_{p,\lambda}} \qquad \text{Eq 1}$$

In this way unwanted variations due to possible changes in lighting over the time can be minimised. The reflectance data obtained were stored in three-dimensional arrays, usually called hypercubes, where the rows and columns are the image dimensions (expressed in pixels), while the third dimension describes the spectral wavelengths along the whole NIR range (a total of 270 variables).

In order to process the hypercubes by means of multivariate data analysis, it was necessary to preliminarily unfold each 3D array to obtain a 2D matrix where the rows are the pixels of the image (row times columns of the hypercubes) and the columns are the spectral channels [18]. Subsequently, the pixels related to the background (area of the image not covered by the sample) and to the cheese holes were removed thorough a two-step masking process based on the wavelength ratio (WR) approach. In more detail, the ratio between the reflectance at two selected pairs of wavelengths was calculated, for enhancing the spectral differences – and the contrast in the resulting image – between the cheese and the background (WR1) and, consecutively, between the cheese and the cheese holes (WR2). The value of the total intensity after ratio calculation was used for setting a threshold that defines the pixels belonging to the cheese in respect to the ones belonging to background or holes. Equation 2 reports the calculation of the first wavelength ratio (WR1). The corresponding threshold for excluding background pixels was chosen by evaluating the image histogram at WR1 (not shown): if the intensity of WR1 was higher than 0.3, the pixel was excluded from the sample (and from the subsequent data analysis).

$$WR_1 = \frac{R_{1455\,nm}}{R_{1305\,nm}} \qquad \text{Eq 2}$$

Equation 3 reports the calculation of the second wavelength ratio (WR2). In this case, the corresponding threshold for excluding hole pixels, evaluated as described above, was set as follows: if the intensity of WR2 was lower than 0.45, the pixel was excluded from the sample (and from the subsequent data analysis).

$$WR_2 = \frac{R_{1260\,nm}}{R_{1050\,nm}} \qquad \text{Eq 3}$$

In each image, the intensities of the spectra retained after background and holes elimination were normalised by a range-scaling between 0 and 1 (Malegori et al., 2020). This normalization permitted to compare, in a reliable way, images acquired in different analytical sessions.

From a spectral point of view, the unfolded hypercubes were pre-treated using, as the row pre-processing, the standard normal variate (SNV) transform, effective to correct both the baseline shift and the global intensity changes in the set of signals [19].

After the image preparation described, which was in common for all the cheeses investigated, a different data processing strategy was followed according to the two steps presented. This separation will be maintained in the following sections, for the sake of clarity.

STEP A – biochemical process understanding: from the hyperspectral images acquired for each of the twelve cheese typologies considered, two square regions of interest (ROIs) of 20x20 pixels were selected. All the pixels belonging to the ROIs and related to the same cheese were considered as a representative characterisation of sample heterogeneity. The ROIs were selected avoiding the inclusion of unwanted interferences such as cheese crust, big holes and evident moulds on the surface, with the aim of focusing on cheese bulk. With the aim of understanding the spectral differences between the cheese typologies, a global matrix was built through the combination of the ROIs – after the spectral pre-processing and image normalization described – and submitted to an exploratory PCA with column mean centering. A joint interpretation of scores and loadings was performed, aimed at identifying the spectral bands ascribable to variations in water, protein and lipid content.

STEP B – biochemical mapping over time: for the twenty hyperspectral images of independent Formaggetta cheeses, a pixel-based approach was followed (Malegori, Grassi, Marques, Freitas, & Casiraghi, 2016), with the aim of visualising the spatial extent of the biochemical phenomena in false-colour maps. According to this aim, three separated PCA (with column mean centering) were carried out on the three NIR regions selected in STEP A (one for each biochemical phenomenon under investigation), after the concatenation of the 10 images acquired over time for both transversal and longitudinal sections. The scores of the lowest-order principal component (PC1) calculated for each single pixel were then represented as an image, thanks to the refolding strategy; the images were coloured according to a colour scale ranging between blue (lowest value of PC1) and red (highest value of PC1).

Reasearch outcomes

STEP A – biochemical process understanding

For an easier reading of the figures, a colour code was used: in grey, the cheeses characterised by a reduced maturation; in pink, those with a marked proteolytic maturation and, in blue, those with a typical lipolytic ripening (Atlante sensoriale dei prodotti alimentari, 2012). Moreover, each of the 12 commercial cheese samples was coded with a number: 1 = Avetino; 2 = Morbidezza; 3 = Bel Paese; 4 = Formaggetta; 5 = Asiago; 6 = Casera; 7 = Camoscio D'Oro; 8 = President; 9 = Gorgonzola; 10 = Roquefort; 11 = Parmigiano Reggiano 22 months; 12 = Parmigiano Reggiano 30 months.

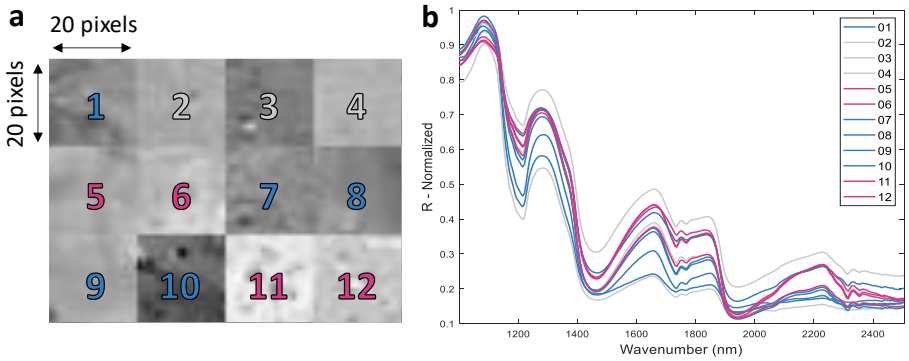

Figure 1: Mean reflectance spectrum from ROIs of cheese samples.

The spatial concatenation of the ROIs (one for each cheese, as an example) is represented in Fig.1a; the grey lightness, which differs visually from ROI to ROI after normalization, is proportional to the sum of the reflectance (R) values along the whole NIR range (total-intensity image, or integral image). Fig.1b shows the mean normalised spectrum along the NIR spectral range for each commercial cheese, obtained as the mean of the reflectance values for all the pixels belonging to the same sample. Strong differences can be highlighted between the mean spectra along the whole spectral range and, in particular, between the three cheese typologies. In more detail, the proteolytic cheeses (in pink) are characterised by a higher absorption between 1600-1800 nm; a more confused and not stable spectral profile can be highlighted from 1900 nm until the end of the recorded NIR region. Conversely, for the light blue spectra (lipolytic cheeses), this ending region (around 2200 nm) seems to be more characteristic, exhibiting a similar behaviour in the whole set of lipolytic cheeses. On the full set of spectra (one spectrum for each pixel), PCA was performed in order to understand which spectral variables are more accountable for sample groupings, according to the typical ripening process. In Figure 2, the outcomes of PCA for the two lowest-order principal components (PCs), which globally explain about 92% of the total variance, are presented.

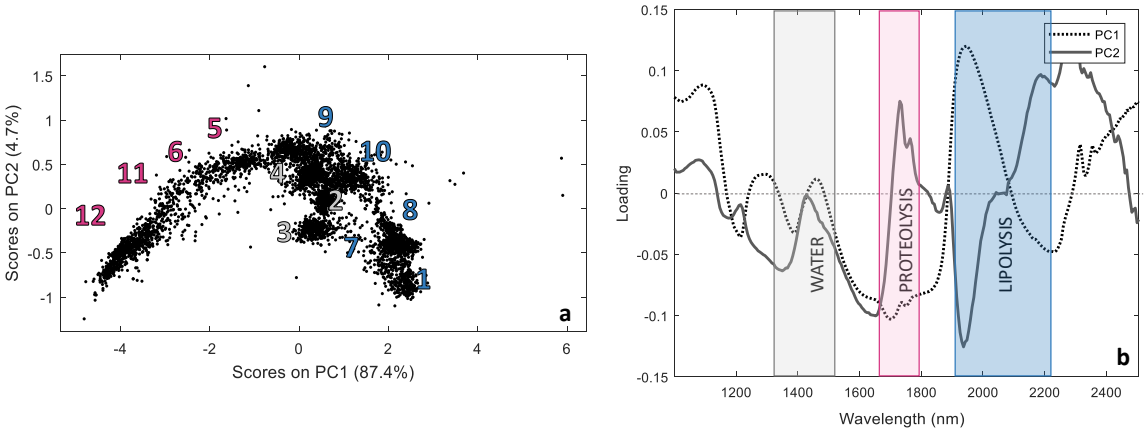

*Figure 2:(a) PC1-PC2 score plot of the concatenated 12 ROIs, one for each cheese (20*20 pixels). In (b) PC1-PC2 loading line plot.*

In the score scatter plot (Fig.2a), where each ROI pixel is represented by a single dot, it is possible to highlight a cloud of objects, centered in the axes origin and distributed along PC1. In more detail, samples characterised by a higher proteolytic activity, corresponding to numbers 5, 6, 11 and 12, are placed at negative values of PC1 scores while, at higher score values, samples 1, 7, 8, 9 and 10 – which correspond to the cheese varieties with a marked lipolytic activity – are found. Around the zero values for PC1 scores, at the origin of the axes, samples 2, 3 and 4 are located, which present a lower extent of maturation and an intermediate behaviour for the two biochemical phenomena of interest. From these considerations, it is possible to conclude that PC1 is explaining the different ripeness type, but without the possibility of deconvolving the contribution of proteolysis and lipolysis phenomena.

In the loading line plot (Fig. 2b), the variable contribution for PC1 and PC2 is reported as a function of the wavelength. From a joint interpretation of sample groupings and variable importance, it is possible to highlight three significant bands: 1360-1500 nm (highlighted in grey), with a modest contribution for both PC1 and PC2; the second one, between 1650-1780 nm (in pink), characterised by a negative correlation with PC1 and a positive correlation with PC2; the last one, around 1920-2205 nm (in light blue), with an opposite behaviour with respect to the previous one, characterised by a positive correlation with PC1 and a negative one with PC2. Based on these outcomes,

it is possible to relate the band around 1690 nm to the proteolytic phenomenon, ascribable to the samples located at negative values of PC1; conversely, samples located at positive PC1 have a positive contribution of the band around 2140 nm and, consequently, they are ascribable to the lipolytic phenomenon.

For proceeding with spectral band imputation, as a confirmation of what PCA is highlighting, it is crucial to underline that the data were analysed in the reflectance mode; consequently, relationships between scores and loadings must be interpreted in a reverse way, considering that an increasing in reflectance values corresponds to a decrease in the absorption due to a peculiar chemical bond [20]. Specifically, the spectral region around 1690 nm can be assigned to the CONH2 group, characteristic for a specific secondary structure of proteins (β-sheet): this band – presenting negative loading values on PC1 – is directly linked, in PCA, to the proteolytic cheeses, which are found at negative scores along PC1. This indicates, for these samples, a decrease in the intensity of the NIR absorption by the peptide bond, due to the degradation of proteins that occurs in cheeses characterised by a strong proteolytic ripening. Regarding the lipolysis process, the band around 2140 nm is related to the second overtone of C=O stretching, which is located in the centre of the spectral range between 1920 nm ad 2205 nm and it may be ascribable to the ester bond of lipids. This spectral region – showing positive loadings on PC1 – is directly related with lipolytic cheeses, which are found at positive PC1 scores. This indicates that, in these samples, the NIR absorption due to the ester bonds is lower than in the other samples. Such an outcome can be attributed to the hydrolysis of ester bonds and to a partial oxidation of C=C double bonds. Finally, loadings of PC1 and PC2 close to zero in the region between 1360–1500 nm – related to combination bands of symmetric and antisymmetric O–H stretching [21] – indicate that the contribution of water is negligible in the PCA outcomes reported and studied in Figure 2.

STEP B – biochemical mapping over time:

For visualising the spatial and temporal extent of the three biochemical phenomena in Formaggetta cheese, independent PCA were performed on the three spectral regions identified in STEP A (one for each biochemical process under investigation).

According to the pixel-based approach [17], false colour images were then refolded for representing the score value of PC1 at each pixel. Below, the three maps are reported, organised as follows: on the left part of the figure (identified with A) the score images for the cheeses cut along the longitudinal section are reported, while the transversal cut is reported in the right part (identified with B). For both of the sections, 10 sub-images are presented, one for each independent Formaggetta cheese analysed daily for the whole ripening process, from t1 (first day of analysis) until t10 (last day of analysis). The colour bar ranges from pure blue, minimum value of PC1 score, to pure red, maximum value of PC1 score, passing through green, which represents an intermediate value.

In Figure 3, the chemical map for the dehydration process is reported as a result for the PCA performed on the reduced spectral range 1360-1500 nm.

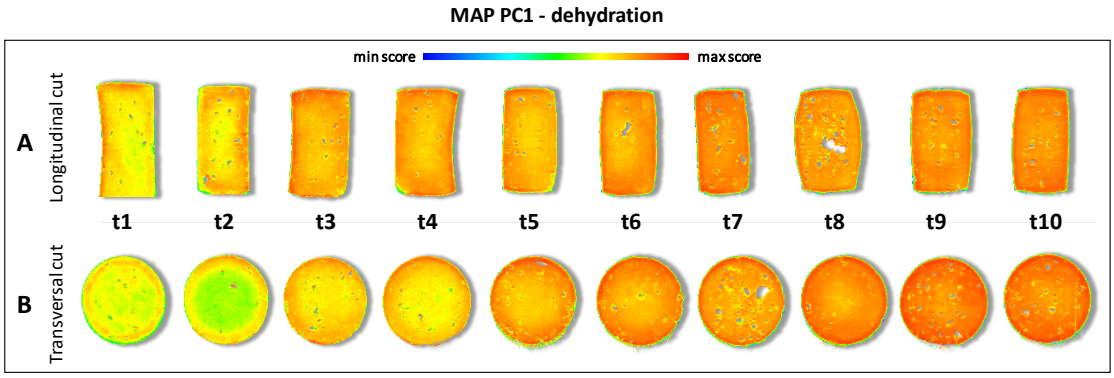

Figure 3: PC1 score map for dehydration

It is possible to notice the profound changes related to the water content in the first five sampling times (from t1 to t5): gradually, the green-yellowish color, associated with lower values of PC1 scores, changes to orange; not an evident spatial pattern is highlightable, except for the transversal cut, where a ring pattern is detectable in t1 and t2. The process tends to reach a plateau at t6, corresponding to the sixth day of ripening, when it is possible to detect the formation of the cheese rind. So, it is possible to conclude that the dehydration process is predominant in the first stage of the

ripening but, after few days, the protection that the rind gives to the product, is counteracting the loss of water.

In Figure 4, the chemical map for the proteolysis process is reported as a result for the PCA performed on the reduced spectral range from 1650 to 1780 nm.

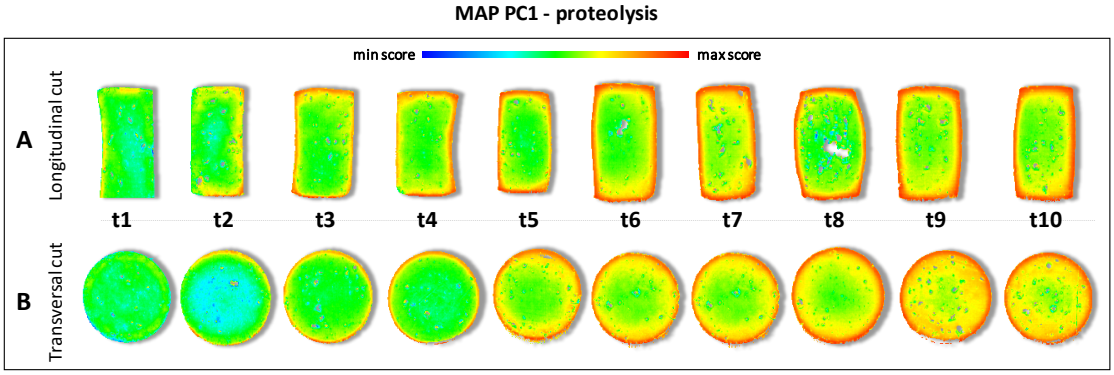

Figure 4: PC1 score maps for proteolysis.

With respect to Figure 3, an opposite situation can be highlighted. In fact, the orange/red colour, limited to the cheese surface in the early sampling time, becomes more pronounced for the last sampling point and, in particular, starting from t5. Moreover, a yellow halo close to the border becomes more and more evident over the ripening stages, suggesting a radial path of the proteolytic phenomena, from the rind to the centre of the cheese.

In Figure 5, the chemical map for the lipolysis process is reported as a result for the PCA performed on the reduced spectral range 1920–2205 nm.

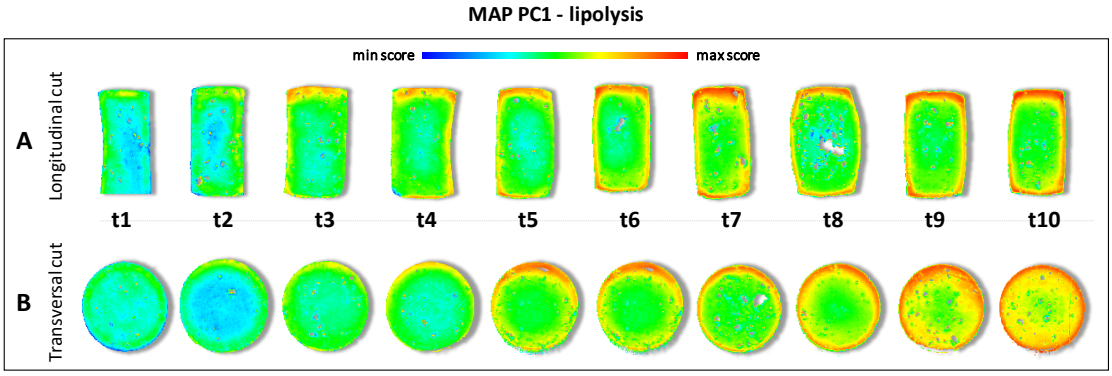

Figure 5: PC1 score map for lipolysis.

Compared to the other two processes investigated, this phenomenon is less evident, starting from blueish pixels in the first sampling time and reaching an intermediate level (predominance of green pixels) at t7/t8. It is possible to appreciate modifications related to the lipolysis process just at the last two sampling times (t9 and t10) with a yellow area and a red border similar to the proteolysis maps. Although in 10 days a reduced lipolytic ripening is highlighted, it is possible to confirm that lipolysis follows a radial path, proceeding from the rind to the centre.

Thanks to the normalisation strategy followed during image pre-processing, the chemical maps are fully comparable; in this way, a direct comparison of the relative importance of the biochemical phenomena in the ripening of the cheese under study can be done. On these bases, Formaggetta can be defined as a semi-hard cheese mainly interested by a proteolytic ripening, which is the most evident phenomenon after rind formation.

Eventually, for all the biochemical processes, no significant differences were highlighted in the spatial extent of the phenomena between longitudinal and transversal sections.

Conclusion and scientific impacts

In light of the presented results, it is possible to confirm the successful applicability of HSI-NIR for understanding the spatial and temporal extent of the three most important biochemical processes that occur during cheese ripening. In more detail, the pixel-

based approach applied for the image processing permitted to visualise, in a straightforward way, the chemical modifications inside the samples, thanks to the representation of the PCA scores as false colour images. This information is of the upmost interest, both for cheese manufacturers and associations involved in the protection of the certified products (consortia) to dispose of a fast, non-invasive and sensitive method to monitor dynamic transformations during maturation in relation to cheese quality. From a general perspective, the present work demonstrated the potential of HSI-NIR, coupled with a chemometric strategy of pattern recognition, as a powerful analytical tool for providing a global understanding of biochemical phenomena involved in heterogeneous matrices. In more detail, the advantage of chemical mapping is demonstrated to be crucial in extracting the information of the spatial extent of chemical changes.

Thanks to this case study, a reliable analytical approach for deeply understanding biochemical changes in food matrices in general is proposed. It is well known, in fact, how the biochemical processes (from ripening to fermentation) can be considered as a primary aspect, to guarantee quality and safety of a food product, together with the development of desired sensorial properties. As a future perspective, given the non-destructive nature of HSI-NIR measurement, a potential application of the method developed in the present study can be foreseen in the early identification of defects related to abnormal ripening processes.

[1] Amigo, J. M., Martí, I., & Gowen, A. (2013). Hyperspectral imaging and chemometrics: a perfect combination for the analysis of food structure, composition and quality. In Data Handling in Science and Technology. Vol 28 (pp. 343-370).

[2] Dorrepaal, R., Malegori, C., & Gowen, A. (2016). Tutorial: Time series hyperspectral image analysis. Journal of Near Infrared Spectroscopy, 24(2), 89-107).

[3] Malegori, C., & Oliveri, P. (2018). Principal component analysis. In *Hyperspectral Imaging Analysis and Applications for Food Quality* (pp. 85-107). CRC Press.

[4] Osborne, B. G. (2006). Near-infrared Spectroscopy in Food Analysis. Near-infrared Spectroscopy in Food Analysis. Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation., 1–14. https://doi.org/10.1002/9780470027318.a1018

[5] Cevoli, C., Gori, A., Nocetti, M., Cuibus, L., Caboni, M. F., & Fabbri, A. (2013). FT-NIR and FT-MIR spectroscopy to discriminate competitors, non-compliance and compliance grated Parmigiano Reggiano cheese. Food Research International, 52(1), 214–220. https://doi.org/10.1016/j.foodres.2013.03.016

[6] Cattaneo, T. M. P., Giardina, C., Sinelli, N., Riva, M., & Giangiacomo, R. (2005). Application of FT-NIR and FT-IR spectroscopy to study the shelf-life of Crescenza cheese. International Dairy Journal, 15(6–9), 693–700. https://doi.org/10.1016/j.idairyj.2004.07.026

[7] Karoui, R., Pillonel, L., Schaller, E., Bosset, J. O., & De Baerdemaeker, J. (2007). Prediction of sensory attributes of European Emmental cheese using near-infrared spectroscopy: A feasibility study. Food Chemistry, 101(3), 1121–1129. https://doi.org/10.1016/j.foodchem.2006.03.012

[8] Currò, S., Manuelian, C. L., Penasa, M., Cassandro, M., & De Marchi, M. (2017). Technical note: Feasibility of near infrared transmittance spectroscopy to predict cheese ripeness. Journal of Dairy Science, 100(11), 8759–8763. https://doi.org/10.3168/jds.2017-13001

[9] Soto-Barajas, M. C., González-Martín, M. I., Salvador-Esteban, J., Hernández-Hierro, J. M., Moreno-Rodilla, V., Vivar-Quintana, A. M., … Curto-Diego, B. (2013). Prediction of the type of milk and degree of ripening in cheeses by means of artificial neural networks with data concerning fatty acids and near infrared spectroscopy. Talanta, 116, 50–55. https://doi.org/10.1016/j.talanta.2013.04.043

[10] Gowen, A. A., O'Donnell, C. P., Cullen, P. J., Downey, G., & Frias, J. M. (2007). Hyperspectral imaging - an emerging process analytical tool for food quality and safety control. Trends in Food Science and Technology, 18(12), 590–598. https://doi.org/10.1016/j.tifs.2007.06.001

[11] Dorrepaal, R., Malegori, C., & Gowen, A. (2016). Tutorial: Time series hyperspectral image analysis. Journal of Near Infrared Spectroscopy, 24(2), 89–107. https://doi.org/10.1255/jnirs.1208

[12] Burger, J., & Geladi, P. (2006). Hyperspectral NIR imaging for calibration and prediction: A comparison between image and spectrometer data for studying organic and biological samples. Analyst, 131(10), 1152–1160. https://doi.org/10.1039/b605386f

[13] Calvini, R., Michelini, S., Pizzamiglio, V., Foca, G., & Ulrici, A. (2020). Exploring the potential of NIR hyperspectral imaging for automated quantification of rind amount in grated Parmigiano Reggiano cheese. Food Control, 112(November 2019), 107111. https://doi.org/10.1016/j.foodcont.2020.107111

[14] Priyashantha, H., Höjer, A., Saedén, K. H., Lundh, Å., Johansson, M., Bernes, G., … Hetta, M. (2020). Use of near-infrared hyperspectral (NIR-HS) imaging to visualize and model the maturity of long-ripening hard cheeses. Journal of Food Engineering, 264, 109687. https://doi.org/10.1016/j.jfoodeng.2019.109687

[15] Ardö, Y., McSweeney, P. L. H., Magboul, A. A. A., Upadhyay, V. K., & Fox, P. F. (2017). Biochemistry of Cheese Ripening: Proteolysis. Cheese: Chemistry, Physics and Microbiology: Fourth Edition, 1(2), 445–482. https://doi.org/10.1016/B978-0-12-417012-4.00018-1

[16] Fox, P. F. (1989). Proteolysis During Cheese Manufacture and Ripening. Journal of Dairy Science, 72(6), 1379–1400. https://doi.org/10.3168/jds.S0022-0302(89)79246-8

[17] Oliveri, P., & Malegori, C. (2019). Principal component analysis. In M. K. N.C. Basantia, L.M.L. Nollet (Ed.), Hyperspectral Imaging Anal. Appl. Food Qual. FL, USA: CRC Press, Boca Raton. https://doi.org/10.1201/b17700-1

[18] Oliveri, P., Malegori, C., Casale, M., Tartacca, E., & Salvatori, G. (2019). An innovative multivariate strategy for HSI-NIR images to automatically detect defects in green coffee. Talanta, 199 (February), 270–276. https://doi.org/10.1016/j.talanta.2019.02.049

[19] Oliveri, P., Malegori, C., Simonetti, R., & Casale, M. (2019). The impact of signal pre-processing on the final interpretation of analytical outcomes – A tutorial. Analytica Chimica Acta, 1058, 9–17. https://doi.org/10.1016/j.aca.2018.10.055

[20] Malegori, C., Alladio, E., Oliveri, P., Manis, C., Vincenti, M., Garofano, P., … Berti, A. (2020). Identification of invisible biological traces in forensic evidence by hyperspectral NIR imaging combined with chemometrics. Talanta, 215(December 2019), 120911. https://doi.org/10.1016/j.talanta.2020.120911

[21] J. Workman, L. Weyer (2007) Practical Guide to Interpretive Near-Infrared Spectroscopy, 10.1002/anie.200885575J. Workman, L. Weyer (2007) Practical Guide to Interpretive Near-Infrared Spectroscopy, 10.1002/anie.200885575

5. OVERALL CONCLUSION

The vibrational spectroscopic techniques investigated in this Doctoral Thesis have proved to be one of the most efficient and advanced tools for developing innovative analytical protocols for quality control in food and pharmaceutical industries.

To meet the emerging analytical challenges in these fields, it is crucial to define alternative methods for performing fast, non-destructive and accurate analysis for monitoring manufacturing processes and verifying the quality of final products. In respect to traditional absorption spectroscopic techniques, Excitation-Emission Fluorescence spectroscopy coupled with three-way decomposition methods such as PARAFAC, proved to have higher sensitivity and specificity for the measurements of low concentrated targets compounds in aqueous solutions. The potential of this approach allowed to obtain satisfying results for characterizing antioxidant compounds in green tea samples with the aim to provide an alternative analytical tool for the authentication of the tea samples according to the geographical origin. The high performance ensured by fluorescence spectroscopy also open the possibility to develop innovative diagnostic protocols based on the analysis of biological fluids for the early detection of prostate cancer. The preliminary study reported in this thesis has shown the potential of Excitation-Emission fluorescence spectroscopy coupled with multivariate data analysis for discriminating patients affected by prostate cancer and healthy donors through the analysis of urine samples with a great improvement of the patient compliance.

In the present thesis, the second vibrational spectroscopic technique tested was Near Infrared Spectroscopy which proved to be a very powerful tool for both lab and industrial applications. NIR benchtop instruments have been used for the analysis of finished food products as olive oil and rice germ. These systems provided very accurate results for the development of predictive regression models for the quantification of key quality parameters and for understanding complex processes that occur in biological matrices during storage. The availability of more sophisticated techniques, such as hyperspectral imaging, which evaluate also the spatial information

related to the sample, allowed to map, and visualize the evolution of these dynamic chemical processes over the time. In this thesis, a combined approach of HSI and pattern recognition techniques proved to be an efficient tool for following the hydrolysis, lipolysis and proteolysis during cheese ripening demonstrating as the contribution of the spatial information can be crucial to obtain a comprehensive knowledge about complex phenomena as the biochemical ones. The same strategy could be applied to study the shelf-life or ripening processes of many different food matrices in which multiple reactions can affect the sensorial and nutritional properties occur at the same time of the product.

In the last years, due to the introduction of new Quality by Design (QbD) approaches for providing statistical, analytical and risk-management methodologies in the design, development and manufacturing of food and pharmaceutical formulations, the interest in applying NIR sensors directly online for monitoring the evolution of manufacturing processes is significatively grew up. In this thesis, real industrial case-studies allowed to develop and compare different chemometric strategies for a comprehensive understanding of a powder blending process by NIRS. Miniaturized NIR sensors proved to be able to obtain satisfactory analytical performances in acquiring spectral information during the mixing process of multicomponent formulations. The collaboration with Medi Nova, a leader Italian company in animal artificial insemination, allowed to test an ad-hoc qualitative solution for the endpoint detection of the blending of semen extender. The good results obtained through the calculation of a Multivariate Statistical Process Control (MSPC) model have been implemented in a dedicated interface developed in the MATLAB environment and currently used in the plant for controlling the efficiency of the production. Moreover, I had the possibility to deepen this topic during my internship abroad in Barcelona where I investigated and compared different qualitative approaches for defining a promising multi-step strategy for the global understanding of an industrial blending process.

Chapter 5