December
2019

Chemometrics Methods Applied to Non-Selective Signals in Order to Address Mainly Food, Industrial and Environmental Problems

PhD Thesis

Maryam Hooshyari

University of Genova

Doctorate in Sciences and Technologies of Chemistry and Materials

Chemometrics Methods Applied to Non-Selective Signals in Order to Address Mainly Food, Industrial and Environmental Problems

PhD Thesis

Curriculum: SAFC

XXXII Cycle

Maryam Hooshyari

Supervisor: Prof. Monica Casale

Data Sheet

Title: Chemometrics Methods Applied to Non-Selective Signals in Order to Address Mainly Food, Industrial and Environmental Problems
Subtitle: PhD thesis

Author: Maryam Hooshyari
Supervisor: Prof. Monica Casale
Department: Pharmaceutical, Food and Cosmetology Sciences, Research Group of Analytical Chemistry and Chemometrics
Curriculum: Pharmaceutical, Food and Cosmetology Sciences (SAFC)

University: Università degli Studi di Genova

Financial support: Università degli Studi di Genova

Thesis Abstract:

Chemometrics is a chemical discipline that uses mathematical and statistical methods in order to extract useful information from multivariate chemical data. Moreover, chemometrics is applied to correlate quality parameters or physical properties to analytical instrument data such as calculating pH from a measurement of hydrogen ion activity or a Fourier transform interpolation of a spectrum. Aim of this thesis project is to develop chemometrical strategies for the elaboration and the interpretation of non-selective complex data in order to solve real problems in food, industry and environmental fields.

Keywords: Chemometrics, PCA, PARAFAC, SIMCA, PLS-CM, PLS-DA, ANOVA, PLS, QDA, D-Optimal Design, Green Tea, Lichen Thalli, Air Pollution, Engine Oil, Base Oil, Crude Oil, Naphthenic Acid, Produced Water, NIR, Fluorescence, LC- HRMS, UV-Visible, Spectroscopy, Chromatography.

Table of Contents

Preface

During the three years of my PhD, the exceptional possibilities provided by Chemometrics to effectively extract information from multivariate and aspecific (non-selective) data obtained with advanced analytical instruments such as spectrofluorimeter, near infrared (NIR) spectrometer and liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS), have been investigated.

The aim of the present thesis was to find simple answers to different real problems, in particular in the food, environmental and industrial fields, using chemometric tools and strategies in order to elaborate multivariate data, also in integrated way.

The achievements of these projects were five published papers in international scientific journals, two oral presentations and ten posters at international scientific conferences.

The **First Chapter** comprehends three studies in the food field ("Food Projects"). The research activity of "**Project I**" was performed at the University of Burgos, Spain under the supervision of Prof. M. Cruz Ortiz Fernandez; the aim was to optimise, by a D-Optimal design coupled with Parallel Factor Analysis (PARAFAC), signals from molecular fluorescence spectroscopy in order to obtain the best experimental conditions for the achievement of the best fluorescence signal of green tea samples.

The parameters optimized thanks to this investigation were utilised in a second study, "**Project II**", for obtaining information on the geographical origin of green tea, in particular for the characterisation of Chinese and Japanese samples, using excitation–emission matrix (EEM) fluorescence spectroscopy and PARAFAC. Moreover, in this study, a cyclodextrin-modified micellar electrokinetic chromatography method was employed to quantify the most represented catechins and methylxanthines in the green tea samples and Partial Least Squares Class-Modelling (PLS-CM), as a multivariate classification tool, was performed on these electrokinetic chromatography data in order to discriminate tea samples according to their geographical origins. The achievements of this project were outlined in an article and two posters.

In “**Project III**”, the analytical performances of quartz cuvettes and disposable glass vials for the analysis of olive oil by near infrared spectroscopy (NIRS) were considered and compared. This project was supported by AGER Foundation, Project Code: 2016-0169. For this purpose, a set of extra virgin olive oil samples from different Italian olive-growing areas have been collected and analysed using both quartz cuvette and mono-use glass vials. From spectral data, multivariate calibration models were developed to estimate quality parameters of extra virgin olive oil: methyl esters of fatty acids and triacylglycerols determined by a fast-GC approach and an UHPLC system, respectively. Before computing the regression models, an optimisation procedure of spectra pre-treatment was performed in order to individuate the pre-treatment able to properly enhance the information embodies in the data. The predictive ability of each PLS model was evaluated by an external validation procedure with an independent test set. The Passing- Bablok linear regression was lastly used to statistically compare the performances of the two different types of cuvettes. In light of the outcomes of the present study, analytical performance of quartz cuvettes and disposable glass vials were considered not significantly different in predicting the olive oil quality parameters taken into account.

The **Second Chapter** deals with the scientific activities that I carried out in the environmental field during my PhD. My work on this field began with a biomonitoring study ("Project IV"); this study aimed at testing the use of different analytical spectroscopic approaches, coupled with chemometrics, as rapid and simple tools for assessing effects of air pollutants on lichen thalli. For achieving this goal, thalli of the fruticose lichen *Pseudevernia furfuracea* (L.) Zopf v. *furfuracea*, collected from a pristine area, have been transplanted for three months to 15 sites in the Liguria region (NW-Italy), characterized by contrasting levels and type of atmospheric pollution, as measured by the regional Environmental Protection Agency (ARPAL). Lichen samples have been analyzed by FFFS (Front-Face Fluorescence Spectroscopy), NIRS and PEA (Plant Efficiency Analyzer) and data elaborated by multivariate data analysis, in order to compare the performances of these spectroscopic techniques and to highlight possible synergic or complementary information.

The outcome of "**Project V"** as a published article was based on my activities performed in NIVA institute, Oslo (Norway), under the supervision of Dr. Saer Samanipour. In this project, the ability of three different extraction methods (liquid-liquid extraction recommended by Norwegian Oil and Gas for extraction produced water, solid phase extraction using Hydrophilic-Lipophilic-Balanced cartridges, and the combination of ENV+ and C8 cartridges) for separation of naphthenic acids (NAs) in oilfield produced water, was evaluated by analysing the data acquired by liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS). The importance of this project is due to the high toxicity of NAs on most kind of organisms and to the corrosively determined by the structure of the naphthenic acid. For each extraction method, one sample divided in three aliquots was tested. In order to evaluate the performance of the three extraction methods, we performed both uni-variate and multi-variate statistical analysis and our results suggested that different extraction methods have different ability to extract toxins from the same sample.

In **Third Chapter,** details of my studies in the industry field are provided. “**Project VI”** investigates the possibility of determining the base oil type in engine oils by combining excitation-emission fluorescence spectroscopy (EEM), NIR spectroscopy and Chemometrics. The purpose of this project is to significantly reduce the cost and time of engine oil formulators (in particular additive package formulators) and standardizers. To this end, I have collaborated with three petrochemical companies to collect samples with specific information, including fifty-three base stocks and forty-three engine oils with various base oil compositions and different performance levels. The performances of both spectroscopies were compared using chemometrics tools such as: PCA for the visualization of pure base stocks and engine oils and PLS-DA as a classification technique in order to distinguish base stocks according to their API (American Petroleum Institute) category. Considering the 3-ways nature of the EEM data, PARAFAC was also applied on fluorescence data as a 3-ways decomposition method.

Chapter 1

Chemometric Strategies in Food Projects

1.1 Project I

D-Optimal Design and PARAFAC as Useful Tools for the Optimisation of Signals from Fluorescence Spectroscopy Prior to the Characterisation of Green Tea Sample

Summary

A procedure based on a D-optimal design coupled with PARAFAC was proposed to optimise signals from molecular fluorescence spectroscopy to obtain the best experimental conditions for the achievement of the best fluorescence signal of green tea samples. Excitation-emission signals (EEMs) were used to analyse the liquid samples (tea infusions), whereas front-face fluorescence excitation-emission matrices (FFEEMs) were recorded for the solid samples (raw or powder tea leaves). The experimental effort was reduced considerably in both cases thanks to the D-optimal design. Once the optimal conditions have been found, the characterisation of green tea was carried out and the sensitivity and specificity were evaluated. The projection of the principal component analysis (PCA) scores enabled to differentiate among the types of liquid green tea (Chinese tea, Chinese tea with lemon and Indian tea with and without theine). The discrimination of solid green tea according to its geographical origin (Chinese, Indian and Japanese) was also carried out through PCA. In addition, the discrimination between the most expensive Japanese tea and the cheapest one was possible. The sensitivity of the models built with SIMCA was 100% and the specificity of the models for the Chinese tea with respect to the Japanese tea was also high.

1.1.1 Introduction

Tea is a beverage made from the leaves of the *Camellia sinensis* plant which is successfully cultivated and consumed by a wide range of age groups in many different countries [1]. The Asian region has a good reputation in the international market due to the high quality of teas produced [2]. A recent study carried out by the Food and Agriculture Organization of the United Nations (FAO) [3] shows that tea production in the world was 5,063,900 tonnes in 2013, China and India being the main producers. Japanese tea is one of the most valued. The most widely drunk grade of green tea in Japan is called Sencha, whereas the highest quality Japanese green tea is Gyokuro whose price is high [1;4-5].

The intake of green tea has been shown to reduce the risk of cardiovascular disease and certain types of cancer [6]. These health benefits are attributed to the high content of catechins (polyphenolic compounds) which have potent antioxidant functionality [7] and native fluorescence [8]. Tea leaves also contain other chemical constituents such as caffeine, theanine, polyphenols, vitamins, minerals, carbohydrates and pigments [1; 9].

The potential health benefits of green tea, especially its antioxidant properties, have increased its consumption. These characteristics vary according to the region in which tea has been cultivated [10] so the price depends on the geographical origin. Consumers would be willing to pay more for a tea produced in a specific geographical region in which tea is considered of higher quality. Therefore, the recognition of the origin is crucial to protect the interests of consumers and sellers [11]. Several analytical methods have been proposed to characterise the geographical origins and/or varieties of teas [11-15].

Fluorescence spectroscopy is a fast, non-destructive, sensitive and low-cost technique which can be used in food authentication without the use of time-consuming sample preparation. Front-face fluorescence spectroscopy measures the fluorescence emitted from the sample surface and avoids some problems such as inner-filter effect, scattering light that are present, for example, on turbid samples [16-17]. Nitin Seetohul et al. [18] discriminated Sri Lankan black teas using fluorescence spectroscopy and linear discriminant analysis. Dong et al. [19] used a fast light-emitting diode (LED)-based 2D fluorescence correlation spectroscopy technique to predict the quality (price) of tea. EEM fluorescence spectroscopy coupled with the NPLS- DA technique was used to discriminate the variety and grade of liquid green tea after derivatization to obtain a series of amino acid derivatives [20]. A recent study applied partial least squares class modelling (PLS-CM) to the content of catechins and methylxanthines of green tea samples by cyclodextrin-modified micellar electrokinetic chromatography [21].

The interpretation of fluorescence spectral data is complex due to overlapping signals. However, the use of fluorescence spectroscopy coupled with chemometric tools such as PARAFAC enables the estimation of the spectra of the underlying fluorescent phenomena [22]. PARAFAC will be considered as a datadriven deconvolution in this work. Many factors are involved in an analytical procedure which may need to be optimised. In this context, the methodology based on the design of experiments can be used to find the best experimental conditions in an effective way.

In this work, front-face fluorescence spectroscopy was used in the analysis of solid green tea samples, whereas the analysis of liquid samples was carried out using conventional fluorescence spectroscopy. Once the optimal conditions have been found, the characterisation of different varieties of green tea infusions and green tea samples in solid form was performed through principal component analysis (PCA) unfolding the same data into matrices. Different SIMCA models have been built with these scores that enable the evaluation of their sensitivity and specificity. As far as the authors are aware, this is the first time that the characterisation of solid green tea leaves using front-face fluorescence spectroscopy was performed evaluating the sensitivity and specificity of the models built.

1.1.2 Material and Methods

1.1.2.1 Samples and Reagents

Commercial samples of green tea from China, India and Japan were analysed. In particular, two varieties of green tea of the same geographical origin were purchased: Chinese tea, Chinese tea with lemon, Indian tea with theine, Indian tea without theine, Japanese tea (Gyokuro) and Japanese tea (Sencha). A green tea sample of unknown origin was also analysed.

Methanol (CAS no. 67-56-1) (gradient grade for liquid chromatography LiChrosolv®) was obtained from Merck KGaA (Darmstadt, Germany) and used to clean the faces of the window of the powder holder when necessary. All the tea infusions were prepared using deionised water obtained with the Milli-Q gradient A10 water purification system from Millipore (Bedford, MA, USA).

1.1.2.2 Experimental Procedure

The samples analysed came from a single green tea bag or from a mixture of three bags according to the experimental plans contained in Tables 1.1.2 and 1.1.4 (“Optimisation of the Procedure by Means of a D-Optimal Design”), whereas three tea bags were used in “Classification of Green Tea”.

The raw solid samples were crushed with a manual mortar until obtaining a powder. The sample had to be powdered as finely and as homogeneously as possible to avoid surface structure effects. Then, the cell was filled with the sample ensuring a uniform distribution of the sample and the powder holder was finally placed into the front surface accessory.

The liquid samples (tea infusions) were prepared by putting 0.2 g of tea into contact with 10 mL of water at 85 °C for 5 min in a beaker. Then, the beaker was placed into an ice bath for 30 s and the content was filtered before its measurement. A filter paper (Albet® LabScience, 73 g/m^2) was used to prepare the tea infusions.

1.1.2.3 Instrumental

The excitation-emission fluorescence measurements were performed at room temperature on a PerkinElmer LS 50B Luminescence spectrometer (Waltham, MA, USA) equipped with a front surface accessory and a powder holder for the measurement of the solid tea. In the case of the liquid tea, the excitation-emission matrices were recorded using the standard cell holder and a 10-mm quartz SUPRASIL® cell with a cell volume of 3.5 mL by PerkinElmer (Waltham, MA, USA). The excitation spectra were recorded between 200 and 290 nm (each 5 nm), whereas the emission wavelengths ranged from 295 to 550 nm (each 1 nm). The excitation monochromator slit width was set to 10 nm. The emission monochromator slit width was set to 5 or 10 nm and the scan speed was 200, 500 or 1000 nm min−1 depending on the experimental plan.

1.1.2.4 Multivariate Data Analysis

1.1.2.4.1 PARAFAC Decomposition

PARAFAC resembles PCA, but while PCA works on a twodimensional matrix, PARAFAC is able to model n-way data. In the case of three-way data, PARAFAC decomposes a data tensor **X** with dimension $I \times J \times K$ into three loading matrices **A**, **B** and **C**. The columns of these loadings matrices are $\mathbf{a_f}$, $\mathbf{b_f}$ and $\mathbf{c_f}$ respectively [28]. The trilinear PARAFAC model is:

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

where x_{ijk} is the element in the position i, j, k of the three-way tensor **X**; F is the number of factors; and e_{ijk} is the residual tensor.

The excitation-emission fluorescence matrices obtained for several samples can be arranged into a three-way tensor and the PARAFAC decomposition can be applied to the analysis of fluorescent data. Therefore, the vectors $\mathbf{a_f}$, $\mathbf{b_f}$ and $\mathbf{c_f}$ are named as the excitation, emission and sample profiles of the *f*-th fluorophore, respectively. The excitation and emission profiles refer to the excitation and emission spectra of each fluorophore, whereas the sample profile corresponds to the amount of each fluorophore found in each sample.

Data are trilinear when the experimental data tensor is compatible with the structure in Eq. (1). The core consistency diagnostic (CORCONDIA) [28] is an index that measures the degree of trilinearity of the experimental data tensor which should be close to 100%.

If the fluorescence data are trilinear, the PARAFAC decomposition provides unique profile estimations when the appropriate number of factors has been chosen to fit the model. PARAFAC has been widely used due to this highly attractive uniqueness property [29], which could be used for the unequivocal identification of compounds.

1.1.2.4.2 D-Optimal Experimental Design

A D-optimal design [11] can be used to reduce the number of observations substantially without losing efficiency. Furthermore, it is possible that the effect of one or more factors on the response is not linear, so three levels should be considered.

Briefly, the steps followed in the selection of the D-optimal design were:

i) Define the factors to study and their levels establishing all the possible candidate experiments, N_C.

ii) Propose a model and establish the number of its coefficients (p). The mathematical reference-state models considered in this work were the ones in Eqs. (2) and (4), respectively. The minimum number of experiments necessary to fit the model that must be extracted from the complete factorial design is p.

iii) Verify the coherence between the model and the information obtained in the candidate points through the variance inflation factors (VIFs).

iv) Construct several experimental matrices with information of enough quality for different values of the number of experiments, N. N varies between the minimum value possible (p) and a value smaller or equal to N_C.

v) The final number of experiments of the D-optimal design is chosen through an exchange algorithm [11] with a value of VIFs that guarantees precise estimations for the coefficients of the model.

1.1.2.5 Software

The FLWinLab software (PerkinElmer) was used to register the fluorescent signals. The Rayleigh signals were removed using INCA software [23]. PARAFAC and PCA models were performed with the PLS_Toolbox [24] for use with MATLAB [25]. The D-optimal designs were built with NemrodW [26]. SIMCA models were performed with V-PARVUS [27].

1.1.3 Results and Discussion

1.1.3.1 Optimization of the Procedure by Means of a D-Optimal Design

1.1.3.1.1 Solid Green Tea Samples

The optimisation of the procedure for the analysis of the green tea samples in solid form was carried out in two experimental sessions. Therefore, a blocked experimental design was considered in order to study the effect of the experimental session (block) on the response. Table 1.1.1 shows the factors under consideration together with their corresponding levels. The block (factor 1, x1) was studied at two levels: first day (level A) and second day (level B). The aim of the sampling (factor 2, x_2) is to obtain a representative part of the material under study. In this work, each tea bag contained a different amount and length of branches and leaves. So, the samples analysed came from a single tea bag (level A) and from a mixture of three bags (level B). The sample preparation (factor 3, x_3) was also considered. The solid samples were measured in raw form (level A) and in powder form (level B). The emission slit width (factor 4, x_4) is the spectral band width of the emission monochromator. In general, higher fluorescence signals are obtained with a wider slit setting. However, the best spectral resolution is obtained when a narrow slit width is selected. The two levels were: 5 nm (level A) and 10 nm (level B). The time of data acquisition is directly related to the scan speed. An optimal signal-to-noise ratio would be achieved by selecting a slow scanning speed but the analysis will take more time to finish and it would also cause the degradation of photochemically sensitive samples. Therefore, the scan speed (factor 5, x_5) was studied at three levels: 200 nm min^{-1} (level A), 500 nm min^{-1} (level B) and 1000 nm min^{-1} (level C).

Table 1.1.1: Experimental domain for the optimization of the procedure for the solid tea samples

Factor (units)	**Codified Variable**	**Level A**	**Level B**	**Level C**
Block (day)	x_1	1	2	-
Sampling (number of bags mixed)	x_2	1	3	-
Sample preparation	x_3	Raw	Powder	-
Emission slit width (nm)	x_4	5	10	-
Scan speed (nm min^{-1})	x_5	200	500	1000

The full factorial design necessary to handle four factors at two levels and another one at three levels would have $2^4 \times 3^1 = 48$ experiments. The mathematical reference-state model that relates the levels of the factors to the response variable is expressed in Eq. (2).

$$y = \beta_0 + \beta_{1A}x_{1A} + \beta_{2A}x_{2A} + \beta_{3A}x_{3A} + \beta_{4A}x_{4A} + \beta_{5A}x_{5A} + \beta_{5B}x_{5B} + \varepsilon \qquad (2)$$

The discrete variables x_{ij} (i = 1, 2, 3, 4, 5 and j=A,B) of the model of Eq. (2) codify the factor and level according to the values previously mentioned.

The model of Eq. (2) had 7 coefficients, so at least 7 out from the 48 experiments of the full factorial design were necessary to fit the model. In this case, a D-optimal design with 16 experiments was chosen (see Table 1.1.2). The values of the VIFs ranged from 1 to 1.22 which indicate high precision in the coefficients of the fitted models.

Chinese green tea samples were prepared and measured according to the experimental plan (Table 1.1.2). The whole experimental procedure is detailed in “Experimental Procedure”. The FFEEMs recorded for the 8 experiments of the first block were arranged in a data tensor, whereas another tensor was built with the FFEEM data of the experiments of the second block.

Table 1.1.2: Experimental plan and response (PARAFAC sample loadings) of the D-optimal design selected for the optimization of the procedure for the solid tea samples.

Experiment	Block (day)	Sampling (number of bags mixed)	Sample preparation	Emission slit width (nm)	Scan speed (nm min^{-1})	Response (PARAFAC sample loadings)
1	1	1	Raw	5	200	0.05
2	1	1	Powder	5	200	0.05
3	1	3	Raw	10	200	0.63
4	1	3	Powder	10	200	0.76
5	1	3	Powder	5	500	0.01
6	1	1	Raw	10	500	0.07
7	1	3	Raw	5	1000	0.02
8	1	1	Powder	10	1000	0.08
9	2	1	Raw	5	200	0.04
10	2	1	Powder	5	200	0.04
11	2	3	Raw	10	200	0.33
12	2	3	Powder	10	200	0.87
13	2	3	Powder	5	500	0.03
14	2	1	Raw	10	500	0.33
15	2	3	Raw	5	1000	0.04
16	2	1	Powder	10	1000	0.17

The dimension of both tensors was 19 × 256 × 8, where the first and second dimensions corresponded to the number of excitation and emission wavelengths recorded, respectively, and the third dimension was the number of samples.

A two-factor PARAFAC model was estimated in both cases after a non-negativity constraint had been laid down on the spectral modes. Both factors might correspond to two different groups of fluorophores contained in the Chinese tea analysed. These PARAFAC models had CORCONDIA indexes equal to 100% and 99%, respectively. Fig. 1.1.1 shows the loadings of the excitation, emission and sample profiles of both models. The explained variance of these models was 97.4 and 96.8%, respectively.

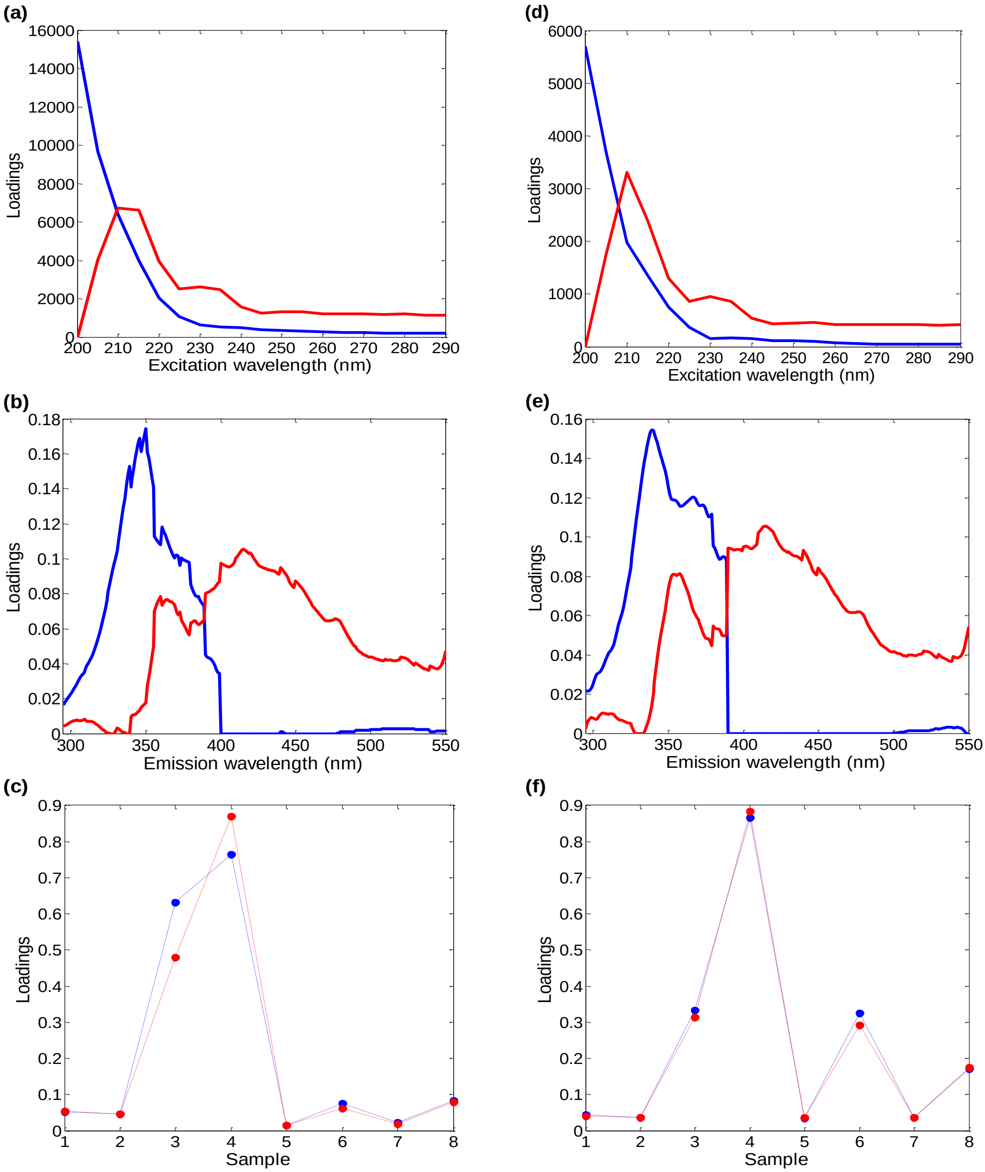

Figure 1.1.1: Two-factor PARAFAC models obtained for the D-optimal design for the solid tea samples. Loadings of the: excitation profile (**a**) and (**d**), emission profile (**b**) and (**e**), and sample profile (**c**) and (**f**). Factor 1: blue, factor 2: red.

As can be seen in Fig.1.1.1 c and f, the sample loadings for the first factor were similar to the ones for the second factor so the experimental conditions had the same influence on both factors increasing or decreasing the fluorescence signal. The sample loadings of the first factor for each experiment (see the last column of Table 1.1.2) were the response considered in the D-optimal design since the explained variance with this factor was higher.

The model for the D-optimal design of Eq. (2) was significant at a 95% confidence level (p value was lower than 0.05) since the null hypothesis is that the model is not significant. The explained variance of the response was 85.1%.

The interpretation of the effect of the experimental factors will be easier if the model of Eq. (2) is converted into the equivalent presence-absence model of Eq. (3) [11]:

$$y = \beta'_0 + \beta'_{1A}x_{1A} + \beta'_{1B}x_{1B} + \beta'_{2A}x_{2A} + \beta'_{2B}x_{2B} + \beta'_{3A}x_{3A} + \beta'_{3B}x_{3B} + \beta'_{4A}x_{4A} + \beta'_{4B}x_{4B} + \beta'_{5A}x_{5A} + \beta'_{5B}x_{5B} + \beta'_{5C}x_{5C} + \varepsilon \qquad (3)$$

where each indicator variable xij is equal to 1 if the factor i-th is at level j-th and zero in any other case. As a consequence, each coefficient β'ij of Eq. (3) is the effect of factor i, at the corresponding level j, on the response. β'ij is a quantity that is added to the response when the factor i is at level j. The estimated coefficients of the presence absence model and their significance are shown graphically in Fig. 1.1.2a. The boundaries of the critical region of the test for the significance of every coefficient at a 95% confidence level are represented in Fig. 1.1.2a as dash-dotted lines. The coefficients placed on the right are positive.

As can be seen in Fig. 1.1.2a, factor 1 (block) was not significant so there was no problem in measuring the tea samples in different days. On the other hand, although factor 3 was not significant, visible differences were observed between the raw and powder samples. The positive coefficient that increased the response was chosen, so the sample will be prepared in powder form for future analyses.

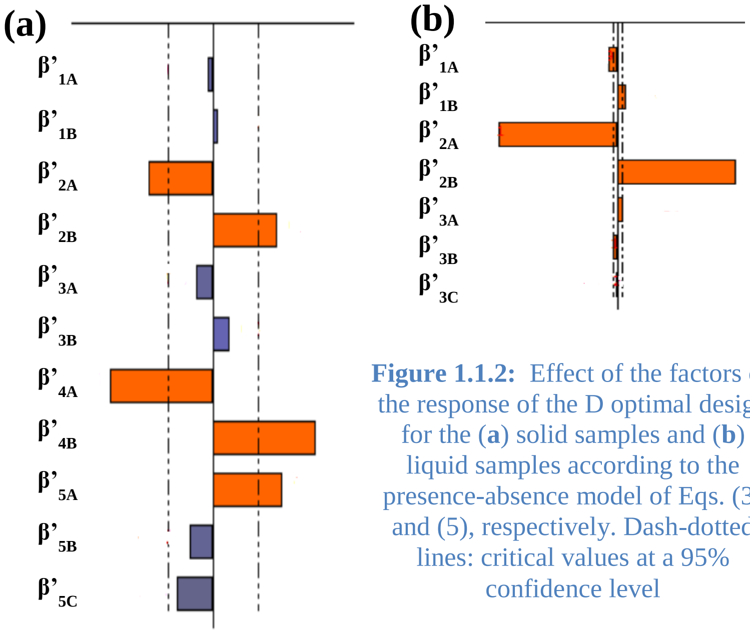

Figure 1.1.2: Effect of the factors on the response of the D optimal design for the (**a**) solid samples and (**b**) liquid samples according to the presence-absence model of Eqs. (3) and (5), respectively. Dash-dotted lines: critical values at a 95% confidence level

The rest of the factors were significant and, considering that a maximum was wanted for the response, the optimal conditions were a mixture of three tea bags (level B), emission slit width at 10 nm (level B) and speed equal to 200 nm min^{-1} (level A).

1.1.3.1.2 Green Tea Infusions

Three experimental factors were considered in the optimisation of the procedure for the liquid tea samples: (i) sampling (factor 1, x_1), (ii) emission slit width (factor 2, x_2) and (iii) scan speed (factor 3, x_3). The first two factors were at two levels, whereas the third one was at three levels as in “Solid Green Tea Samples”. Therefore, the full factorial design would have $2^2 \times 3^1 = 12$ experiments. The block was not considered as a factor in this optimisation since it is known from previous experience that the recording of EEM on different days does not have an effect on the results. Table 1.1.3 shows the levels of the studied factors. In this case, the reference-state model fitted was:

$$y = \beta_0 + \beta_{1A}x_{1A} + \beta_{2A}x_{2A} + \beta_{3A}x_{3A} + \beta_{3B}x_{3B} + \varepsilon \qquad (4)$$

Table 1.1.3: Experimental domain for the optimization of the procedure for the green tea infusions.

Factor (units)	Codified Variable	Level A	Level B	Level C
Sampling (number of bags mixed)	x_1	1	3	-
Emission slit width (nm)	x_2	5	10	-
Scan speed (nm min^{-1})	x_3	200	500	1000

The experimental plan followed in this case for the 8 experiments selected is included in Table 1.1.4. This D-optimal design had values of the VIFs between 1 and 1.22, which guaranteed high precision in the estimation of the coefficients.

The tea infusions were prepared (see “Experimental Procedure”) using Chinese green tea and the samples were measured according to the experimental plan of the D-optimal design (see Table 1.1.4). The Rayleigh signals were not deleted in these data because INCA software (Andersson 1998) deleted these signals together with a significant part of the fluorescence signals of the fluorophores. A three-way data tensor containing the EEM recorded for the 8 experiments was built. However, the EEM corresponding to experiment number 4 of the D-optimal design was considered as an outlier. In addition, some excitation and emission wavelengths were deleted due to high noise. The dimension of the resultant tensor was 12 × 200 × 7. A two-factor PARAFAC model was estimated (CORCONDIA of 93%, explained variance of 71%) after a non-negativity constraint on both spectral ways was imposed. As in “Solid Green Tea Samples”, the sample loadings of the first factor were the response used in the D-optimal design since they were nearly the same as the ones of the second factor. These values are collected in the last column of Table 1.1.4.

Table 1.1.4: Experimental plan and response (PARAFAC sample loadings) of the D-optimal design selected for the optimization of the procedure for the liquid tea samples.

Experiment	Sampling (number of bags mixed)	Emission slit width (nm)	Scan speed (nm min^{-1})	Response (PARAFAC sample loadings)
1	1	5	200	0.14
2	3	10	200	0.51
3	1	5	500	0.11
4 [i]	3	5	500	-
5	1	10	500	0.47
6	3	10	500	0.49
7	3	5	1000	0.14
8	1	10	1000	0.47

[i] Outlier

The model of Eq. (4) had five coefficients and 7 experimental results were available so there were enough degrees of freedom to evaluate the significance of the model and the coefficients. The model was significant at a 95% confidence level and the explained variance of the response was 100%.

The reference-state model of Eq. (4) was converted into the presence-absence model of Eq. (5):

$$y = \beta_0 + \beta_{1A}x_{1A} + \beta_{1B}x_{1B} + \beta_{2A}x_{2A} + \beta_{2B}x_{2B} + \beta_{3A}x_{3A} + \beta_{3B}x_{3B} + \beta_{3C}x_{3C} + \varepsilon \qquad (5)$$

Figure 1.1.2b shows the graphic study of the effect of the experimental factors on the response considered in this D-optimal design. In this case, all the factors were significant, but the emission slit width was the most important one. A maximum response was wanted, so the optimal conditions were those in which the significant factors had a positive coefficient, that is a mixture of three tea bags, emission slit width at 10 nm and speed equal to

200 nm min^{-1}. These conditions were the same as the optimal conditions found in “Solid Green Tea Samples” for the solid tea samples.

1.1.3.2 Classification of Green Tea

1.1.3.2.1 PCA Model for Liquid Green Tea (Chinese and Indian)

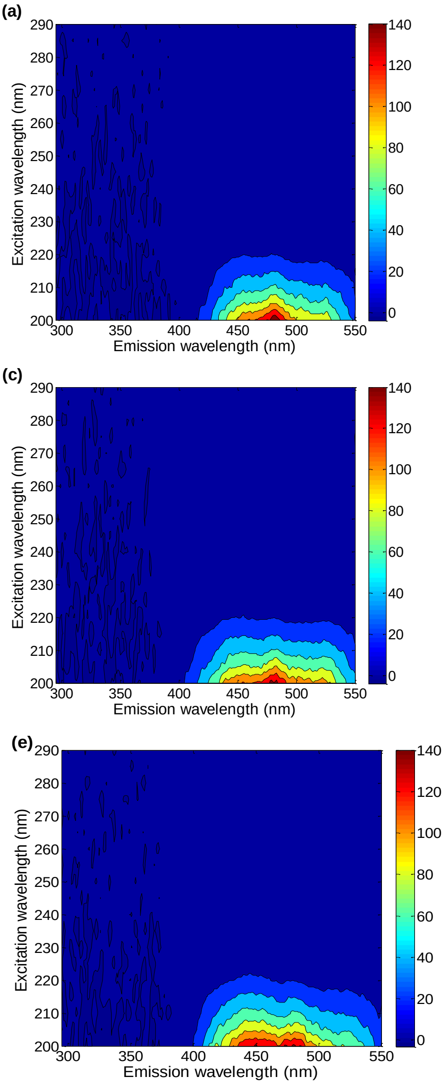

Figure 1.1.3: Contour plots of the EEM recorded of **a** liquid green tea sample: a Chinese, **b** Chinese with lemon, **c** unknown origin, **d** Indian without theine and e Indian with theine

Several tea infusions were prepared using different types of green tea from China and India and measured under the optimal conditions selected in "Green Tea Infusions". These liquid samples were specifically: 3 Chinese tea infusions, 2 Chinese tea infusions with lemon, 3 Indian tea infusions without theine, 3 Indian tea infusions with theine and 2 infusions prepared with green tea from an unknown origin. As in "Green Tea Infusions", the Rayleigh signals were not deleted in these data. Figure 1.1.3 shows the contour plots obtained with the EEM recorded of a liquid sample prepared with each one of the different teas considered.

The EEMs recorded for the 13 samples analysed were arranged into a data matrix of dimension 13 × 3288 since some of the variables had a lot of missing values. PCA was applied on this matrix after the data had been mean-centred. The cross-validation step was carried out by means of the "leave-one-out" technique where the minimum of the cross-validation variance was obtained with two principal components (PCs). The first and second PCs explained 81.4% and 12.0% of the variance, respectively. The

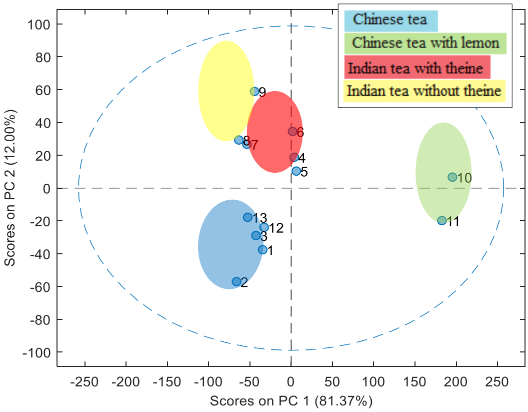

Figure 1.1.4: Score plot of the liquid green tea samples from China and India on the first and second principal components. Samples 12 and 13 correspond to green tea from an unknown origin

representation of the scores for each liquid sample on the first and second PCs is shown in Fig. 1.1.4.

As can be seen in this figure, PCA enabled to differentiate the infusions of green tea according to the four different varieties. The green tea of unknown origin (samples 12 and 13 in Fig. 1.1.4) might be from China due to the closeness to this group in that figure. The shape of the contour plots obtained for the samples analysed was quite similar as can be seen in Fig. 1.1.3, so it is clear the difficulty of the analysis. However, the PCA decomposition has succeeded in the characterisation of these tea samples since the scores are in non-overlapped regions.

1.1.3.2.2 PCA Model for Solid Green Tea (Chinese, Japanese and Indian)

Different varieties of green tea from three different origins were used to prepare solid tea samples in powder form (“Experimental Procedure”). These samples were measured under the optimal conditions of “Solid Green Tea Samples”. The number of these solid samples was: 3 Chinese green tea, 3 Chinese green tea with lemon, 2 Indian green tea without theine, 2 Indian green tea with theine, 7 Japanese green tea (Sencha) and 6 Japanese green tea (Gyokuro). An example of the contour plots of the FFEEM obtained in each case is shown in Fig. 1.1.5. These plots are clearly different from the ones obtained for the tea infusions (see Fig. 1.1.3).

In a first step, a data matrix of dimension 23 × 3288 was considered and decomposed by PCA as in “PCA Model for Liquid Green Tea (Chinese and Indian)”. Three PCs were considered which explained a 94.3% of the variance. The score for each solid sample on the first and third PCs is represented in Fig. 1.1.6.

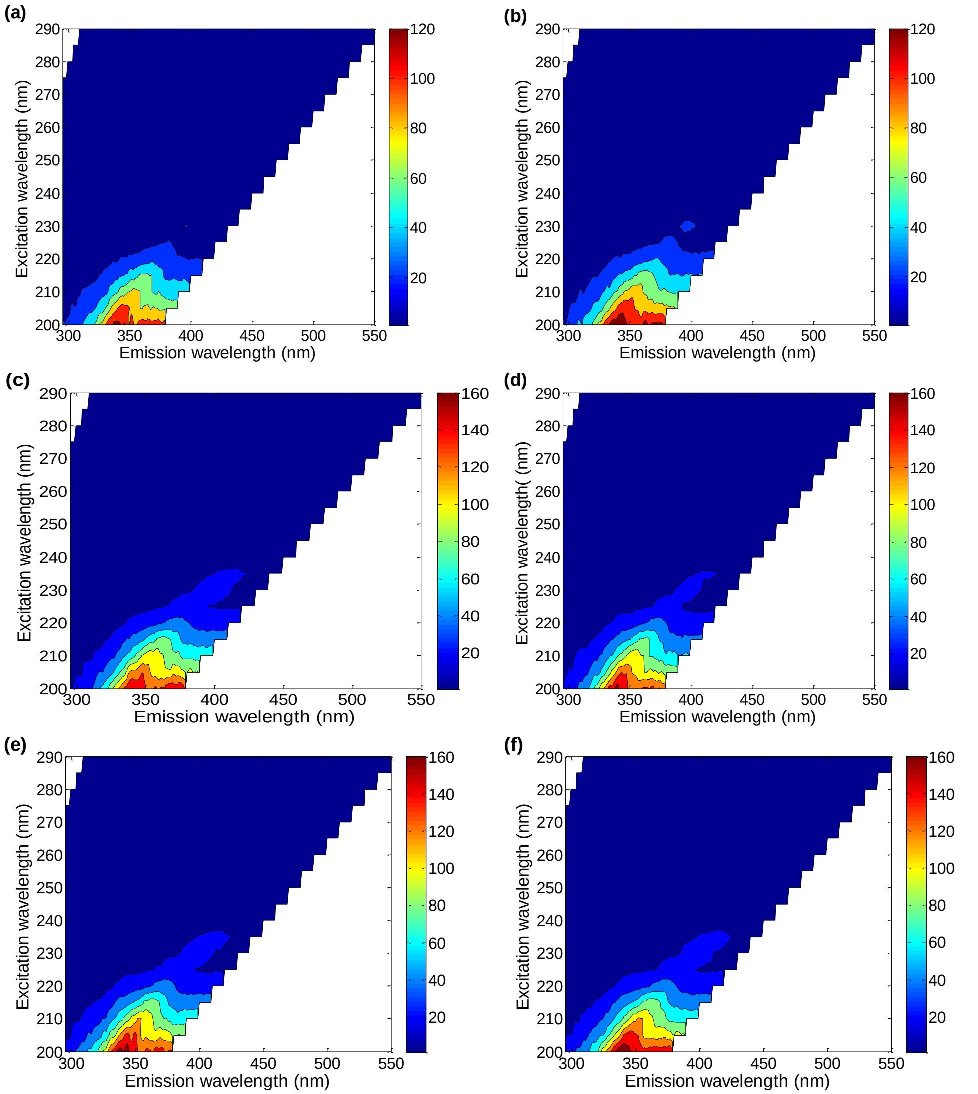

Figure 1.1.5: Contour plots of the FFEEM recorded of a solid green tea sample of: **a** Chinese, **b** Chinese with lemon, **c** Indian without theine, **d** Indian with theine, **e** Japanese (Sencha) and **f** Japanese (Gyokuro)

This plot showed a good distinction between the Chinese, Indian and Japanese solid tea samples. Only samples number 6 and 7 were outliers since these samples came from China and Japan, respectively. Therefore, it

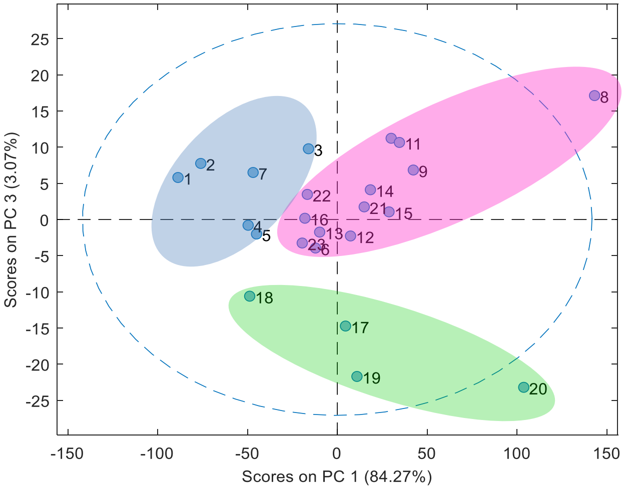

Figure 1.1.6: Score plot of the solid green tea samples on the first and third principal components. Chinese tea (samples 1 to 6): blue, Japanese tea (samples 7 to 16 and 21 to 23): pink and Indian tea (samples 17 to 20): green. Samples number 6 and 7 are outliers.

was possible to discriminate the samples according to their geographical origin despite their spectra being quite similar as can be seen in Fig. 1.1.5.

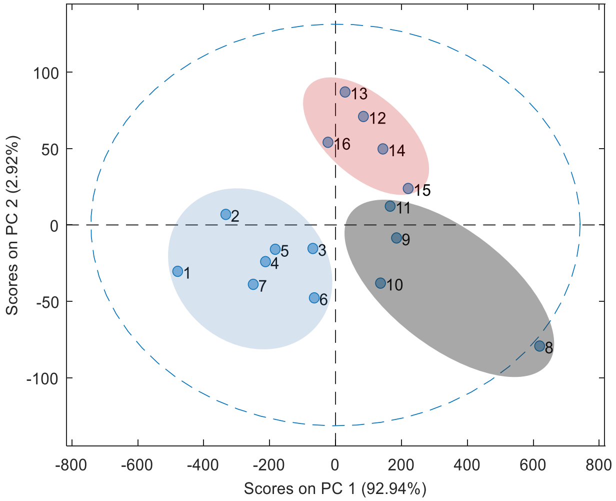

Figure 1.1.7: Score plot of the solid green tea samples on the first and second principal components. Chinese tea (samples 1 to 6): blue, Japanese tea (Sencha, samples 7 to 11): grey and Japanese tea (Gyokuro, samples 12 to 16): pink. Outlier: sample 7

In a second step, only the Chinese and Japanese solid samples were considered in the study. In the case of the Japanese samples, 5 samples were only considered for each variety (Sencha and Gyokuro). A data matrix with these 16 samples was built as explained above. This matrix (16 × 3288) was mean-centred and decomposed by PCA. The “leave-one-out” technique was used to perform the cross-validation step. Two PCs were necessary in this PCA model (explained variance of 95.9%). The scores of the samples on the first and second PC are displayed in Fig. 1.1.7. Only one of the Japanese samples (sample number 7) that corresponded to the Sencha variety was an outlier. The first component enabled the distinction of the samples by their geographical origin (China and Japan), whereas the most expensive Japanese tea (Gyokuro) and the cheapest one (Sencha) could be distinguished with the second PC. The price of Gyokuro doubles the one of Sencha (80 and 40 euros/Kg, respectively in our market).

The differences between the two varieties of Japanese tea are lower than the differences among the tea samples from other origins. Therefore, the Japanese tea could not be discriminated according to the variety in the first analysis. However, when only Chinese and Japanese samples were considered, the discrimination between the two varieties of Japanese tea was possible with a PCA model. This is due to the way PCs are found.

The decision rule to discriminate a new tea sample would be: first, project the signal into the previous PCA analysis, and if the scores correspond to Japanese tea, then the signals would be projected into the second PCA analysis to discriminate the Japanese tea by their price (this procedure is a highlevel data fusion [30]).

1.1.3.2.3 SIMCA Model for Solid Green Tea

A class modelling using SIMCA has been carried out to evaluate the sensitivity and specificity of the model. The sensitivity is the ability of the class models to recognise its own objects whereas the specificitymeasures the capacity of the model to reject objects that do not belong to the class.

It is not possible to perform a class modelling using SIMCA by cross-validation for the four categories of liquid green tea ("PCA Model for Liquid Green Tea (Chinese and Indian)") because each category only contains two or three objects. However, the categories of Chinese, Japanese and Indian tea have been modelled with the data of "PCA Model for Solid Green Tea (Chinese, Japanese and Indian)".

Table 1.1.5: Sensitivity (%) and specificity (%), in fitting and in cross-validation, for the three categories of solid tea

Category (Number of samples)	**Number of PC of SIMCA model**	**Sensitivity (%)**	**Sensitivity CV^a (%)**	**Specificity (%)**	**Specificity CV^a (%)**
Cat1 - Chinese (6)	2	100	83.3	cat1-cat2 = 90.9 cat1-cat3 = 75.0	cat1-cat2 = 82 cat1-cat3 = 75
Cat2 - Japanese (11)	2	100	90.9	cat2-cat1 = 100 cat2-cat3 = 100	cat2-cat1 = 92 cat2-cat3 = 100
Cat3 - Indian (4)	1	100	100	cat3-cat1 = 50.0 cat3-cat2 = 63.0	cat3-cat1 = 46 cat3-cat2 = 58

Four PCs were needed to explain at least a 95% of the variance of the categories. These PCs have been obtained with the unfolded data matrices of "PCA Model for Solid Green Tea (Chinese, Japanese and Indian)". The SIMCA models were built with the scores of the samples in those components using normal range and unweighted augmented SIMCA distance [27]. The objects 7 and 8 were considered as outliers and were not included in this class modelling. The "leave-one-out" technique was used to perform the cross-validation step.

Table 1.1.5 shows the sensitivity and specificity of the models built for the three categories (Chinese, Japanese and Indian) whereas Table 1.1.6 shows the results for the two categories of Japanese tea (Sencha and Gyokuro) of different price.

Table 1.1.6: Sensitivity (%) and specificity (%), in fitting and in cross-validation, for the two varieties of Japanese tea

Category	Number of PC of SIMCA model	Sensibility (%)	Sensitivity CV^a (%)	Specificity (%)	Specificity CV^a (%)
Cat1 – japanese (Sencha)	1	100	80	cat1- cat2 = 100	cat1- cat2 = 90
Cat2 – japanese (Gyokuro)	1	100	80	cat2- cat1 = 60	cat2- cat1 = 62

The similar percentages in fitting and in cross-validation indicate that all the models are stable with a sensitivity of 100% in all of them. The specificity was good except for category 3 (Indian tea) and for the category of Gyokuro (Japanese tea).

PCA analysis is just descriptive. On the other hand, the SIMCA model defines a region for each category which is evaluated through sensitivity and specificity. The procedure to classify a new sample is similar to the one explained in “PCA Model for Solid Green Tea (Chinese, Japanese and Indian)” for the PCA analysis. First, the SIMCA model is applied to that new sample for the three categories. If that sample is assigned to the Japanese category, then the second model is applied.

1.1.4 Conclusions

A fast method based on PCA together with the use of fluorescence spectroscopy has enabled the discrimination of different varieties of green tea. The use of a D-optimal design together with PARAFAC has reduced the experimental effort in the optimization prior to the characterisation of the tea samples.

The procedure followed could be considered a fast and promising method for discriminating green tea by its geographical origin as the current market situation and the increment in the tea trade require. In addition, two different varieties of Japanese tea have been distinguished by their price.

The SIMCA models built show a sensitivity of 100% and high specificity. In addition, they are stable when the cross-validation is performed except for the specificity for category 3 (Indian tea) with the other two categories (Chinese and Japanese tea).

1.1.5 References

1. HaraY (2001) Green tea: health benefits and applications.Marcel Dekker Inc., New York
2. Hicks A (2009) Current status and future development of global tea production and tea products. AU JT 12:251–264
3. FAO (2015) Word tea production and trade. Current and future development, Rome
4. Balentine DA, Harbowy ME, Graham HN (1998) Caffeine. In: Spiller GA (ed) Tea: the plant and itsmanufacture; chemistry and consumption of the beverage. CRC Press, USA, pp 35–72
5. Kapoor MP, Rao TP, Okubo T, Juneja LR (2013) Green tea: history, processing techniques, principles, traditions, features, and attractions. In: Juneja LR, Kapoor MP, OkuboT RT (eds) Green tea polyphenols nutraceuticals of modern life. CRC Press, USA, pp 1–18
6. Saeed M, Naveed M, ArifM, Ullah KakarM,Manzoor R, Ezzat Abd El- Hack M, AlagawanyM, Tiwari R, Khandia R,Munjal A, Karthik K, Dhama K, IqbalHMN, DadarM, Sun C (2017) Green tea (Camellia sinensis) and L-theanine: Medicinal values and beneficial applications in humans-a comprehensive review. Biomed Pharmacother 95:1260–1275
7. Carloni P, Tiano L, Padella L, Bacchetti T, Customu C, Kay A,Damiani E (2013) Antioxidant activity of white, green and black tea obtained from the same tea cultivar. Food Res Int 53:900–908

8. Arts ICW, Hollman PCH (1998) Optimization of a quantitative method for the determination of catechins in fruits and legumes. J Agric Food Chem 46:5156–5162
9. Jagan Mohan Rao L, Ramalakshmi R (2011) Recent trends in soft beverages. Woodhead Publishing India PVT LTD, India
10. Zhang C, Chieh Suen CL, Yang C, Young Quek S (2018) Antioxidant capacity and major polyphenol composition of teas as affected by geographical location, plantation elevation and leaf grade. Food Chem 244:109–119
11. Lewis GA, Mathieu D, Phan-Tan-Luu R (1999) Pharmaceutical and experimental designs. Marcel Dekker, New York Ma G, Zhang Y, Zhang J,Wang G, Chen L, Zhang M, Liu T, Liu X, 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
12. Gonçalves Dias Diniz PH, Ferreira BarbosaM, Tavares de MeloMilanez KD, Pistonesi MF, Ugulino de Araújo MC (2016) Using UV–Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup. Food Chem 192:374–379
13. Palacios-Morillo A, Alcázar A, de Pablos F, Marcos Jurado J (2013) Differentiation of tea varieties using UV–Vis spectra and pattern recognition techniques. Spectrochim Acta, Part A 103:79–83
14. Ye NS (2012) A Minireview of analytical methods for the geographical origin analysis of teas (Camellia sinensis). Crit Rev Food Sci Nutr 52(9):775–780
15. 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 Anal Methods 5:856–860
16. Cabrera-Bañegil M, Hurtado-Sánchez MC, Galeano-Díaz T, Durán-Merás I (2017) Front-face fluorescence spectroscopy combined with second-order multivariate algorithms for the quantification of polyphenols in red wine samples. Food Chem 220:168–176
17. Karoui R, Blecker C (2011) Fluorescence spectroscopy measurement for quality assessment of food systems-a review. Food Bioprocess Technol 4:364–386
18. Nitin Seetohul L, Scott SM, O'Hare WT, Alic Z, Islam M (2013) Discrimination of Sri Lankan black teas using fluorescence spectroscopy and linear discriminant analysis. Sci Food Agric 93:2308– 2314
19. Dong Y, Lu H, Yong Z, Yan C, He S (2015) Fast two-dimensional fluorescence correlation spectroscopy technique for tea quality detection. App Optics 54:7032–7036

20. Hu L, Yin C (2017) Development of a new three-dimensional fluorescence spectroscopy method coupling with multilinear pattern recognition to discriminate the variety and grade of green tea. Food Anal Methods 10:2281–2292

21. Casale M, Pasquini B, Hooshyari M, Orlandini S, Mustorgi E, Malegori C, Turrini F, Ortiz MC, Sarabia LA, Furlanetto S (2018) Combining excitation emission matrix fluorescence spectroscopy, parallel factor analysis, cyclodextrin-modified micellar electrokinetic chromatography and partial least squares class-modelling for green tea characterization. J Pharm Biomed Anal 159:311–317.

22. Møller Andersen C, Mortensen G (2008) Fluorescence spectroscopy: a rapid tool for analyzing dairy products. J Agric Food Chem 56:720– 729

23. Andersson CA (1998) INCA 1.41. Department of Food Science, University of Copenhagen, Denmark. Available from: http://www. models.life.ku.dk/inca . Accessed 30 Nov 2018

24. Wise BM, Gallagher NB, Bro R, Shaver JM,WindigW, Koch RS (2015) PLS Toolbox 7.9.5. Eigenvector Research Inc., Wenatchee

25. MATLAB (2014) version 8.4.0.150421 (R2014b). The Mathworks, Inc., Natick

26. Mathieu D, Nony J, Phan-Tan-Luu R (2015) NemrodW, Version 2015. L.P.R.A.I., Marseille

27. Forina M, Lanteri S, Armanino C, Casolino MC, Casale M, Oliveri P (2012) V-PARVUS 2012. An extendable package of programs for explorative data analysis, classification and regression analysis, Dept. of Pharmacy, University of Genoa.

28. Bro R, Kiers HAL (2003) A new efficient method for determining the number of components in PARAFAC models. J Chemom 17:274– 286

29. OrtizMC, Sarabia LA, SánchezMS,HerreroA, Sanllorente S, Reguera C (2015) Usefulness of PARAFAC for the quantification, identification, and description of analytical data. In: Muñoz de la Peña A, Goicoechea HC, Escandar GM, Olivieri AC (eds) Data handlingin science and technology: fundamentals and analytical applications of multiway calibration. Elsevier, Amsterdam

30. Roussel S, Bellon-Maurel B, Roger JM, Grenier P (2003) Fusion of aroma, FT-IR and UV sensor data based on the Bayesian inference. Application to the discrimination of white grape varieties. Chemom Intell Lab Syst 65:209–219

1.2 Project II

Combining Excitation-Emission Matrix Fluorescence Spectroscopy, Parallel Factor Analysis, Cyclodextrin-Modified Micellar Electrokinetic Chromatography and Partial Least Squares Class-Modelling for Green Tea Characterization

Summary

In this study, an alternative analytical approach for analysing and characterizing green tea (GT) samples is proposed, based on the combination of excitation–emission matrix (EEM) fluorescence spectroscopy and multivariate chemometric techniques. The three-dimensional spectra of 63 GT samples were recorded using a Perkin–Elmer LS55 luminescence spectrometer; emission spectra were recorded between 295and 800 nm at excitation wavelength ranging from 200 to 290 nm, with excitation and emission slits both set at 10 nm. The excitation and emission profiles of two factors were obtained using Parallel Factor Analysis (PARAFAC) as a 3-way decomposition method. In this way, for the first time, the spectra of two main fluorophores in green teas have been found. Moreover, a cyclodextrin-modified micellar electrokinetic chromatography method was employed to quantify the most represented catechins and methylxanthines in a subset of 24 GT samples in order to obtain complementary information on the geographical origin of tea. The discrimination ability between the two types of tea has been shown by a Partial Least Squares Class-Modelling performed on the electrokinetic

chromatography data, being the sensitivity and specificity of the class model built for the Japanese GT samples 98.70% and 98.68%, respectively. This comprehensive work demonstrates the capability of the combination of EEM fluorescence spectroscopy and PARAFAC model for characterizing, differentiating and analysing GT samples.

1.2.1 Introduction

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 con-centration 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 [1]. 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 [1–7]. Among the most abundant catechins in GT there are (+)-catechin, ((+)C), (-)-epicatechin (EC), (-)-epigallocatechin (EGC), (-)-epicatechingallate(ECG), (-)-epigallocatechin gallate (EGCG) [8].

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 pro-file of bioactive compounds are manufacturing process (withering, steaming/pan-firing, rolling, oxidation/fermentation and drying) and storage [8–9]. 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 [10–11]. Several analytical methods have been proposed together with chemometric techniques in order to characterize the geographical origins and/or varieties of teas [12–15]. 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. [14] 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 some of us [10], 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 [9]. 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 matrix (EEM) 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[16] after unfolding the data into matrices and through Parallel Factor Analysis (PARAFAC) [17] 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 our previous work [10], 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) 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.

1.2.2 Material and Methods

1.2.2.1 Samples and Reagents

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. Appendix 1 shows the description of the samples and the corresponding assigned code.

Table 1.2.1: GT samples analysed by the CyD-MEKC method: content of catechins and methylxanthines[i].

Sample Code[ii]	Category[iii]	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	16.4	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	0.00	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	0.00	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

[i] The data are expressed as the average content in mg g−1, dry basis (mean of twodeterminations).

[ii] Sample code refers to the assigned code as described in Appendix 1.

[iii] Category 1: Japanese GT samples; category 2: Chinese GT samples.

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 analysed using the CyD-MEKC method for the quantitation of catechins and methylxanthines (Table 1.2.1).

1.2.2.2 Experimental Procedure

In order to simulate the content of active compounds in a cup of tea, GT samples were prepared by infusion of tea leaves. The samples were prepared immersing 0.2 g of finely powdered tea leaves in 10 mL of water at 85°C for 5 min in a beaker. Then, the beaker containing tea leaves and water was transferred into an ice bath for 30 s to stop the infusion at the same moment for each sample. In order to remove the leaves before performing the analysis, the infusion was filtered using a filter paper (Albet®LabScience) with a porosity equal to 73 g/m^2.

1.2.2.3 Instrumental

1.2.2.3.1 Capillary Electrophoresis

The CyD-MEKC method used for the determination of the com-pounds was derived from a previous study coauthored by one of us [15]. The analyses were carried out using a ^{3D}CE instrumentfrom Agilent Technologies (Waldbronn, Germany) controlled by the software ^{3D}CE 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}[15]. Further information on the CE method and procedure may be found in mentioned Ref. [15].

1.2.2.3.2 Fluorescence Spectroscopy

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.

1.2.2.4 Software

Data analysis was performed in the MATLAB environment [24], 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 [22].

1.2.2.5 Data Analysis

1.2.2.5.1 Data Exploration

PCA [16] 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* × *J* × *K* into three loading matrices **A**, **B** and **C**, being their columns $\mathbf{a_i}$, $\mathbf{b_j}$ and $\mathbf{c_k}$ respectively. The trilinear PARAFAC model is expressed as follows:

$$x_{ijk} = \sum_{f=1}^{F} a_i b_j c_k \quad \mathrm{i} = 1,2,\ldots,\mathrm{I}; \mathrm{J} = 1,2,\ldots,\mathrm{J}\,; k = 1,2,\ldots,K \qquad (1)$$

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 $\mathbf{a_i}$, $\mathbf{b_j}$ and $\mathbf{c_k}$ 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. [19]. Data are trilinear when the experimental data array is compatible with the structure in Eq. (1).

The core consistency diagnostic (CORCONDIA) developed by Bro and Kiers [20] 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 [17]. PARAFAC has been widely used due to this highly attractive uniqueness property [21], which could be used for the unequivocal identification of compounds.

1.2.2.4.2 Variable Selection

The selection of the informative variables was performed by means of SELECT [18], a feature selection technique based on the stepwise decorrelation of the variables, which is implemented in the V-Parvus software [22]. This technique generates a set of decor-related 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 are 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.

1.2.2.4.3 Class Modelling

PLS-CM [23] is a supervised method of classification between two categories (or classes), in our case Japanese or Chinese GT. Itis 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.

1.2.3 Results and Discussion

1.2.3.1 Catechins and Methylxanthines Content

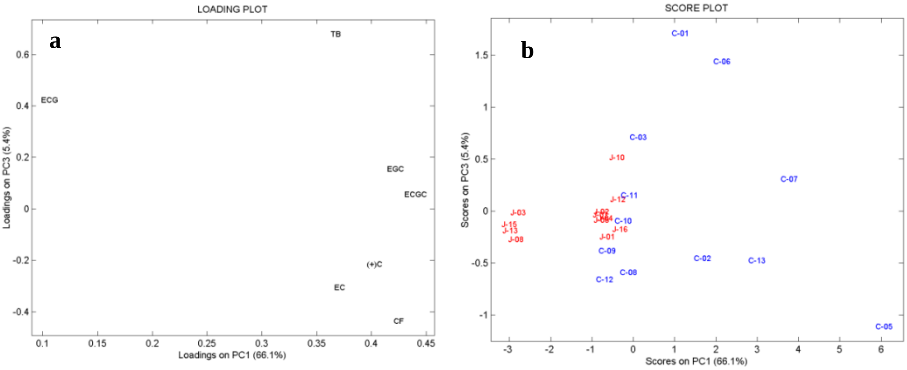

Figure 1.2.1: PCA (**a**) loading plot and (**b**) score plot of catechins and methylxanthines data.

The CyD-MEKC method previously described [15] was applied to the analysis of a subset of 24 GT samples in order to confirm our previous observations [10] 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 1.2.1. This data set was submitted to chemometric modelling starting from PCA as a display method and then applying the PLS-CM algorithm for class modelling 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. 1.2.1 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 [10].

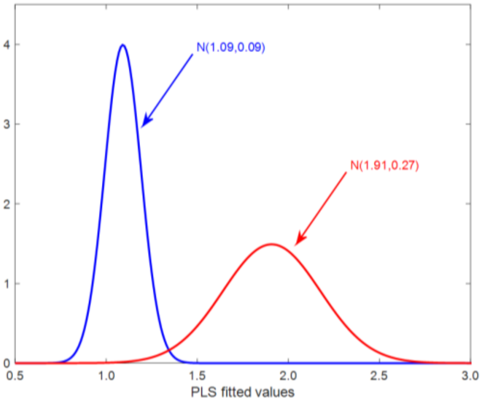

Figure 1.2.2: Normal distribution fitted for Japanese GT samples (on the left) and ChineseGT samples (on the right).

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 1.2.1). 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. 1.2.2 shows the distribution of PLS fitted values for the Japanese and Chinese GT samples. Both classes have normal distribution with mean values 1.09 and1.91 and SD values 0.09 and 0.27, respectively.

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 H0: 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% and98.68%, respectively. The risk curve, reported in Fig. 1.2.3, is the plot of β versus α probabilities, where itis clear that both probabilities change in opposite directions, that is, α decreases when β increases and vice versa.

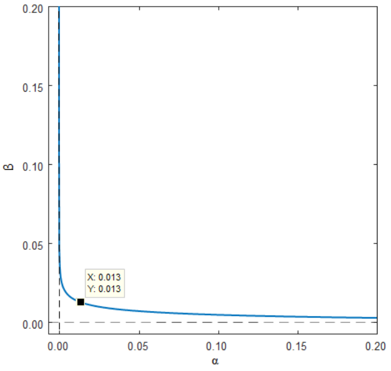

Figure 1.2.3: Risk curve for the PLS-CM.

1.2.3.2 Fluorescence Spectra

Fig. 1.2.4 shows two typical excitation-emission spectra of one Japanese (J1) and one Chinese GT sample (C1).

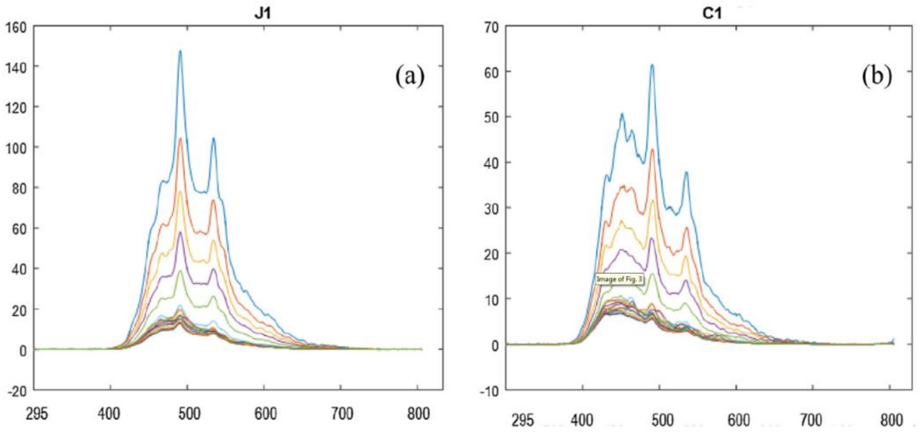

Figure 1.2.4: A typical excitation-emission spectra of (**a**) a Japanese (J1) and (**b**) a Chinese (C1) GT sample.

1.2.3.2.1 Repeatability Studies

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). Supplementary Fig. 1.2.5 displays the score plot obtained by PCA of the spectral data after unfolding. 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.

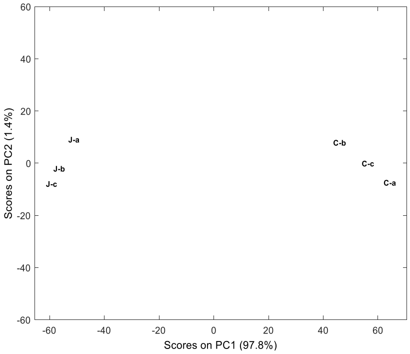

Figure 1.2.5: Score plot obtained by PCA of the spectral data of 3 replicates of 2 GT samples, one Chinese (C5) and one Japanese (J5).

1.2.3.3.2 PCA

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. 1.2.4). 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-centred.

Fig. 1.2.6 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.

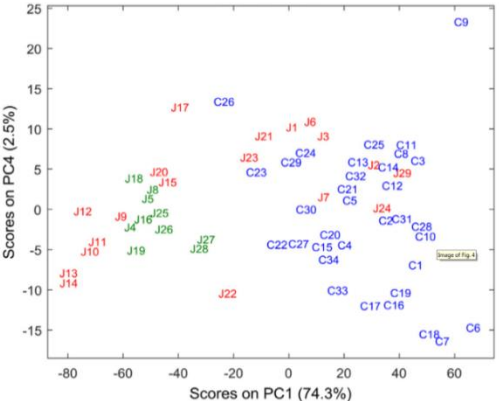

Figure 1.2.6: PCA score plot on the PC1-PC4 plane for the fluorescence data. Matcha samples are indicated in green in the plot (for colours, see the web version of the manuscript).

Looking at the loading profile on PC1 (Fig. 1.2.7), it is possible to notice the bands more informative along PC1 and thus useful for discriminating between Japanese and Chinese GTs, namely 410–450 nm and 500–600 nm. The first band (410–450 nm) shows positive

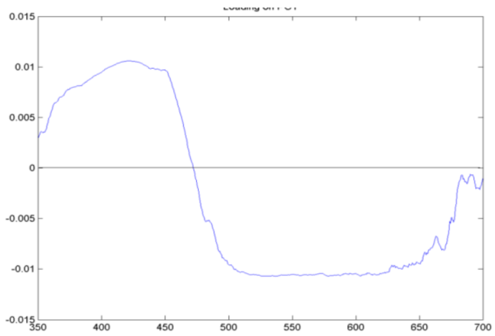

Figure 1.2.7: Loading profile on PC1

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.

1.2.3.3.2 PARAFAC

The EEM data recorded for the 63 samples analysed were arranged into a 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 constrain, required two factors (CORCONDIA of 100%, explained variance of 98.6%).

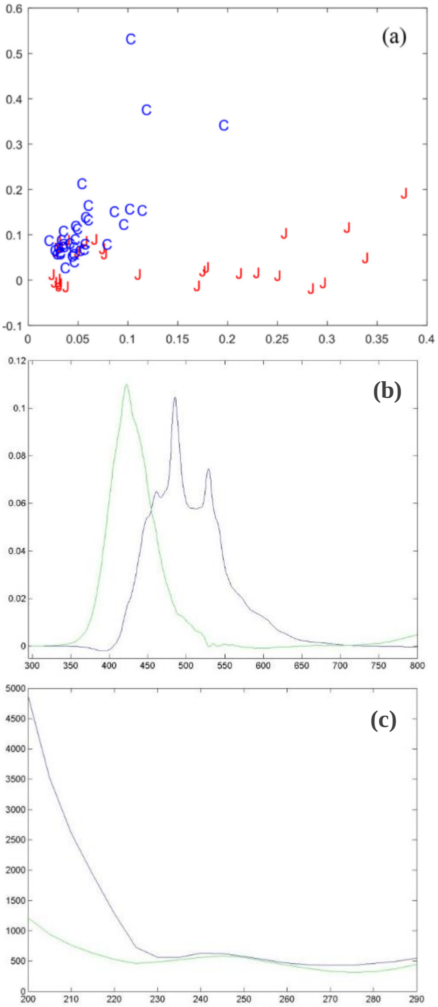

Figure 1.2.8: PARAFAC results: (a) loading plot of the mode of the samples (first mode); explained variance 98.6% (F1 = 96.0% and F2 = 2.6%); (b) loading plot of the emission mode (second mode); (c) loading plot of the excitation mode (third mode).

The plot of the loadings of the mode of the samples (first mode, Fig. 1.2.8a) is similar to the PCA score plot (Fig. 1.2.6) 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. 1.2.8b) shows the emission spectra for two fluorophores, one with maximum around 420 nm and the other one with maxima at 500–550 nm. The plot of the loadings of the third mode (Fig. 1.2.8c) shows the excitation profiles. 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.

1.2.3.3.3 Variable Selection

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 showed as the most selected variables the two bands 415–450 nm and 495–550 nm (Supplementary Fig. 1.2.9).

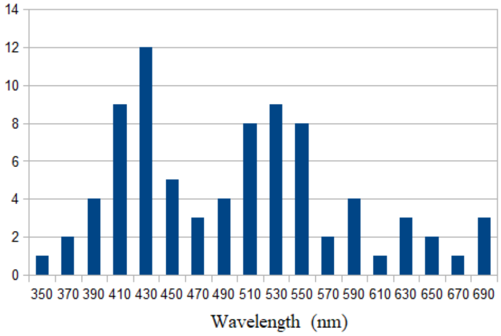

Figure 1.2.9: Variables selected by SELECT: frequency histogram.

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 [25]. The infuses of GT prepared for the analysis were noticed to be slight yellow-green colour due to pigments as chlorophylls and carotenoids; the quantities of pigment extracted in hot water are related to the concentrations of the pigments in teas [26]. These observations were in agreement with the findings of Ref. [27], 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.

1.2.4 Conclusions

The aim of the present study was to evaluate the possibility of using EEM fluorescence spectroscopy as a rapid analytical method for analysing 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 PC1and 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-dimensional emission-excitation 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 wasapplied for the analysis of a subset of 24 GT samples confirming that catechins are more abundant in Chinese samples. In addition, the PLS-CM 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.

Acknowledgements

We would like to acknowledge La Via del Tè (Florence, Italy) for the kind gift of tea samples.

1.2.5 References

1. Y. Suzuki, N. Miyoshi, M. Isemura, Health-promoting effects of green tea, Proc.Jpn. Acad. B-Phys. 88 (2012) 88–101.
2. P. Bogdanski, J. Suliburska, M. Szulinska, M. Stepien, D. Pupek-Musialik, A.Jablecka, Green tea extract reduces blood pressure, inflammatory biomarkers, and oxidative stress and improves parameters associated with insulin resistance in obese, hypertensive patients, Nutr. Res. 32 (2012) 421–427.
3. C. Cabrera, R. Artacho, R. Giménez, Beneficial effects of green tea – a review, J.Am. Coll. Nutr. 25 (2006) 79–99.
4. R. Cooper, Green tea and theanine: health benefits, Int. J. Food Sci. Nutr. 63(2012) 90–97.
5. P. Velayutham, A. Babu, D. Liu, Green tea catechins and cardiovascular health: an update, Curr. Med. Chem. 18 (2008) 1840–1850.
6. H. Wang, G.J. Provan, K. Helliwell, Tea flavonoids: their functions, utilization and analysis, Trends Food Sci. Technol. 11 (2000) 152–160.
7. J.-M. Yuan, C. Sun, L.M. Butler, Tea and cancer prevention: epidemiological studies, Pharmacol. Res. 64 (2011) 123–135.
8. M. Bonoli, P. Colabufalo, M. Pelillo, T. Gallina Toschi, G. Lercker, Fast determination of catechins and xanthines in tea beverages by micellarelectrokinetic chromatography, J. Agric. Food Chem. 51 (2003)1141–1147.
9. A. Kosi´nska, W. Andlauer, Antioxidant capacity of tea: effect of processing andstorage, in: V.R. Preedy (Ed.), Processing and Impact on Antioxidants in Beverages, Academic Press, Elsevier, Waltham, 2014, pp. 109–120.
10. B. Pasquini, S. Orlandini, M. Goodarzi, C. Caprini, R. Gotti, S. Furlanetto, Chiralcyclodextrin-modified micellar electrokinetic chromatography and chemometric techniques for green tea samples origin discrimination, Talanta150 (2016) 7–13.
11. G. Ma, Y. Zhang, J. Zhang, G. Wang, L. Chen, M. Zhang, T. Liu, X. Liu, C. Lu,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 (2016)714–720.
12. P.H. Gonc¸ alves Dias Diniz, M. Ferreira Barbosa, K.D. Tavares de Melo Milanez,M.F. Pistonesi, M.C. Ugulino de Araújo, Using UV–Vis spectroscopy

for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup, Food Chem. 192 (2016) 374–379.

13. N.S. Ye, A minireview of analytical methods for the geographical origin analysis of teas (Camellia sinensis), Crit. Rev. Food Sci. Nutr. 52 (2012)775–780.
14. N. Ye, L. Zhang, X. Gu, Discrimination of green teas from different geographical origins by using HS-SPME/GC–MS and pattern recognition methods, Food Anal. Methods 5 (2012) 856–860.
15. R. Gotti, S. Furlanetto, S. Lanteri, S. Olmo, A. Ragaini, V. Cavrini, Differentiation of green tea samples by chiral CD-MEKC analysis of catechins content, Electrophoresis 30 (2009) 2922–2930.
16. I.T. Joliffe, Principal Component Analysis, Springer-Verlag, New York, 2002.
17. R. Bro, PARAFAC, Tutorial and applications, Chemom. Intell. Lab. Syst. 38(1997) 149–171.
18. M. Forina, S. Lanteri, M. Casale, M.C. Cerrato Oliveros, Stepwis eorthogonalization of predictors in classification and regression techniques: an “old” technique revisited, Chemom. Intell. Lab. Syst. 87 (2007) 252–261.
19. M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, D. Giménez, Identification andquantification of ciprofloxacin in urine through excitation-emission fluorescence and three-way PARAFAC calibration, Anal. Chim. Acta 642 (2009)193–205.
20. R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (2003) 274–286.
21. M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, A. Herrero, S. Sanllorente, C. Reguera,Usefulness of PARAFAC for the quantification, identification, and description of analytical data, in: A. Mu˜noz de la Pe˜na, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Eds.), Data Handling in Science and Technology: Fundamentals and Analytical Applications of Multiway Calibration, Elsevier, Amsterdam,2015, pp. 37–81.
22. M. Forina, S. Lanteri, C. Armanino, M.C. Casolino, M. Casale, P. Oliveri,V-PARVUS, an Extendable Package of Programs for Explorative Data Analysis, Classification and Regression Analysis, Dept. of Pharmacy, University of Genoa, 2014.
23. M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, Tutorial on evaluation of type I and type II errors in chemical analyses: from the analytical detection to authentication of products and process control, Anal. Chim. Acta 674 (2010) 123–142.
24. MATLAB Version 8.4.0.150421 (R2014b), The Mathworks, Inc., Natick, MA,2014.

25. N. Hall, The Tea Industry, first ed., Woodhead Publishing, Cambridge, 2000,pp. 21.
26. Y. Suzuki, Y. Shioi, Identification of chlorophylls and carotenoids in major teas by high-performance liquid chromatography with photodiode array detection, J. Agric. Food Chem. 51 (2003) 5307–5314.
27. M. Lang, F. Stober, H.K. Lichtenthaler, Fluorescence emission spectra of plant leaves and plant constituents, Radiat. Environ. Bioph. 30 (1991) 333–347.

1.3 Project III

PLS Regression Models for the Determination of EVOO Quality Parameters by NIR Spectroscopy: a Comparative Study

Summary

In the present study, the analytical performances of quartz cuvettes and disposable glass vials for the analysis of olive oil by near infrared spectroscopy (NIRS) were considered and compared. Nowadays, laboratories that perform routine analysis on extra virgin olive oil by NIRS employ quartz cuvettes with time-consuming measurements, especially in the washing phase, and an increasing cost to buy and dispose of reagents and to replace eventually damaged cuvettes. The use of mono-use glass vials may reduce times 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 have been collected and analysed using both quartz cuvette and mono-use glass vials. From spectral data multivariate calibration models were developed to estimate quality parameters of extra virgin olive oil: methyl esters of fatty acids (FAMEs) and triacylglycerols (TAGs) determined by a fast-GC approach and an UHPLC system, respectively. Before computing the regression models, an optimisation procedure of spectra pre-treatment was performed in order to individuate the pre-treatment able to properly enhance the information embodies in the data. The predictive ability of each PLS model was evaluated by an external validation procedure with an independent test set. The Passing- Bablok linear regression was lastly used

to statistically compare the performances of the two different types of cuvettes. In light of the outcomes of the present study, analytical performance of quartz cuvettes and disposable glass vials were considered not significantly different in predicting the olive oil quality parameters taken into account.

1.3.1 Introduction

The International Olive Oil Council (IOOC) fixed purity and quality criteria in order to recognize four commercial olive oil categories (or grades): the “extra-virgin” olive oil, the “virgin” olive oil, the “refined olive oil” and the “pomace” [1]. Extra virgin olive oil (EVOO) is considered the highest quality grade and the adulteration with edible oil of inferior quality it’s becoming a type of commercial fraud more and more frequent. The quality criteria established by the IOOC for EVOO include: measurements related to organoleptic characteristics (odour, taste and colour), free acidity, peroxide value, absorbency in ultra-violet at 232 and 270 nm (K 232, K 270, ΔK) and 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 determining olive oils quality [2]. These compounds are considered particularly interesting for their physiological effects [3] and suitable for authenticity assessment of EVOO [4]. In this context, in order 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, has foreseen the employ of innovative analytical protocols, including approaches based on near infrared spectroscopy (NIRS) and multivariate data analysis.

It is well known, in fact, that NIRS nowadays represent a valid and recognise alternative method, compared to traditional techniques, to determine qualitative and quantitative parameters of several food matrices, including olive oil, in a non-destructive way and in few seconds, not requiring sample preparation with a reduction in term of costs and time saving [5]. In literature, in fact, there are several studies that proved the potential of NIRS technology for determining the quality of olive oil both in term of chemical composition [6] and product authentication [7]. Regarding chemical composition, NIRS have been demonstrated to be useful for quantifying important trade standards including peroxide value, free fatty acid content, specific extinction coefficients (e.g. K232 and K270) [7]. Regarding food fraud, NIRS has proven to be an effective analytical method to detect and estimate adulteration of virgin olive oils with vegetable oils of inferior quality [8]. Moreover, in the last decade, NIR spectroscopy has been recognised as an excellent tool for the verification of authenticity of EVOO samples based on their geographical origins [9] or olive cultivar.

The main advantage of NIR technique, is that it is a quick and low-cost method for analysing a large number of samples, but the speed of spectra acquisition can be limited by the employment of quartz cuvettes especially in the washing phase that can include the use of organic solvent, as acetone, with the drawbacks linked to the buying and disposal of these chemicals. In addition, an improper use of these substances also can leave residues in cuvettes with 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. However, due to differences between optical glass vials and quartz cuvettes (QC) in term 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.

In order to fill this gap of knowledge, a comparative study was performed with the aim of understanding if the use of DGV for the NIRS routine 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 QC and DGV. On the obtained spectra, an optimization step of data pre-processing was carried out and then Partial Least Squares (PLS) algorithm [10] was applied on a training set of the NIRS data to estimate the content of FAMEs and TAGs. The prediction ability of these models on a test set of unknown samples was used to compare, for the first time, the analytical performances of the two types of cuvettes. To do this, the Passing-Bablok regression method was applied for performing a joint test on slopes and intercepts of each pairs of models, one using GC and the other one using DGV.

1.3.2 Materials and Methods

1.3.2.1 Samples and Reagents

The 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 are produced with olives harvested in the season 2017-2018. The sampling was planned with the aim of collecting EVOO samples representative of the whole Italian production; in fact, the 106 samples analysed came from ten different Italian regions that represent the most productive areas in the country: Apulia, Tuscany, Sicily, Trentino-South Tyrol, Umbria, Veneto, Calabria, Latium, Sardinia and Liguria. The number of samples analyzed for each region is proportional with the importance of their production (in term of quantity). This set included 28 PDO (Protected Designation of Origin) and 10 PGI (Protected Geographical Indication) EVOO samples; the different olive oil samples were labelled as reported in Appendix 2, where further details about origins, cultivar and designation are given. Thanks to this rational sampling, the national variability of EVOO was taken into account allowing performing a reliable study.

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

1.3.2.2 Experimental Procedure

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 an UHPLC system.

For FAMEs determination, samples were prepared as follows: in a 5 mL screw-top test tube, 25 mg of EVOO sample were weighted. The lipid fraction was transesterified adding 100 µL of the methanolic potassium hydroxide solution (KOH/MeOH, 2M). Thereafter FAMEs were extracted using 1 mL of n-heptane; the reaction mixture was shanked vigorously for 30 seconds. After 5 minutes, the upper FAMEs layer became clear and ready to be injected into GC system. After sample preparation, 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(tripropilfosfonio)dodecano bis(trifluorometilsulfonil) 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 velocity 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). 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.

1.3.2.3 Instrumental

NIR spectra were acquired in trasmittance mode with an FT-NIR spectrophotometer (Buchi NIRFlex N-500, Flawil, Switzerland) in a liquid module 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 controlled temperature (35 ± 0.5 °C)

Samples were acquired in duplicate and the average spectrum for each sample was used for data analysis in order to minimized unwanted spectral variability.

In more detail, EVOO samples were put into a 5 mm pathlength QC directly from the bottle, without any chemical treatment. After the analysis, to prepare the cuvette for further acquisitions, each QC was washed with detergent in warm water, rinsed with 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 GC.

1.3.2.4 Data Analysis

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

First, NIR transmittance spectra were converted into the absorbance scale (Log (1/T)) for a direct interpretability of outcomes [11]. Then, a noisy region at the end of the signal and without significant absorption was removed and the spectral range reduced from 10000 to 4528 cm^{-1}. Subsequently, spectra were organised in two matrices containing 106 rows and 1369 columns, samples and 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 in a training set (including 80% of samples) and a test set (including 20% of samples) thanks to the application of the Kennard and Stone algorithm [12]

Before model computation, a comparison between eight different combinations of data pre-treatments was performed in order to optimize the selection of the most suitable pre-processing strategy and to improve subsequent calibration model. The application of 4 data transformations (two column and two row pre-processing algorithms) was evaluated taking into account not only the application of one transformation at a time but also their combination:

- Mean centring,
- Autoscaling,
- Standard Normal Variate (SNV) + mean centring,
- Orthogonal Signal Correction (OSC) + mean centring,
- SNV + OSC + mean centring,
- SNV + autoscaling,
- OSC + autoscaling,
- SNV + OSC + autoscaling.

SNV was tested, as it allowed to correct baseline vertical shifts and global intensity effects, typically arising from light scattering phenomena in vibrational spectroscopy [11] OSC was evaluated in order to remove some of the information embodied in spectral data that is unrelated (orthogonal) to the qualitative variable to be modelled (Y-vector); in this way just the useful information related to the response is maintained in the X-block [13]. Both the strategies for data normalization (mean centring and autoscaling) were taken into account.

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.

To reach the final aim of statistically comparing the prediction ability of the models built using QC or DGV, the Passing-Bablok regression method [14] was applied on the pairs of Y values predicted by the models developed for each quality parameter separately. The estimation of a linear regression line between two pairs of data column, obtained with two different methods or devices both measured with error, allows to statistically understand the similarity/diversity between the two-independent estimation. To do this, slope and intercept of the fitted line are calculated with their 95% confidence interval. The null hypothesis (H_0) is verified when the slope is not significantly different from 1 and that the intercept is not significantly different from 0.

1.3.3 Result and Discussion

Among the variables describing EVOO quality measured with the reference methods within the Violin project (see previous paragraph 1.3.2.2), six of them, whose range of variability was less restricted than for the other quality parameters, were considered for the comparison between QC and DGV . In more detail, three TAGs and three FAMEs were selected. The TAGs were: dioleoyllinoleoyl-glycerol (OOL), oleoyl-linoleoyl-palmitoylglycerol (OLP) and triolein (OOO), while the FAMEs were: palmitic (C16:0), oleic (C18:1n9) and linoleic (C18:2) acids, that were the most present in the extra virgin olive oil.

Firstly, a subset of 80 EVOO samples was chosen by the Kennard and Stone algorithm (REF) for constituting the calibration set, and the remaining 26 samples were used for the test set, to validate the quality of the regression model in predicting.

In order to select the most suitable strategy to pre-process the NIR spectral profiles, for both QC and DGV data, an optimisation procedure was performed. It is important to underline that independent pre-processing optimizations were performed for QC and DGV data; for each variable considered (three FAMEs and three TAGs) a PLS regression model was computed. Moreover, PLS models were calculated retaining an increasing number of LVs, from 1 until 10, and applying different spectra pre-treatments, according to the list presented in section 2.4. 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, minimises the error of each PLS model in cross-validation. In more detail, Figure 1.3.1 resume 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.

For all the quality parameters modelled using the QC spectra, SNV + OSC + mean centering (represented in green in Figure 1.3.1) turned out to be the best combination, as it allowed to reduce RMSECV with as few LVs as possible. From a global evaluation of the QC models, from 4 to 6 LVs were considered the best compromise between model complexity and associated error (data not shown).

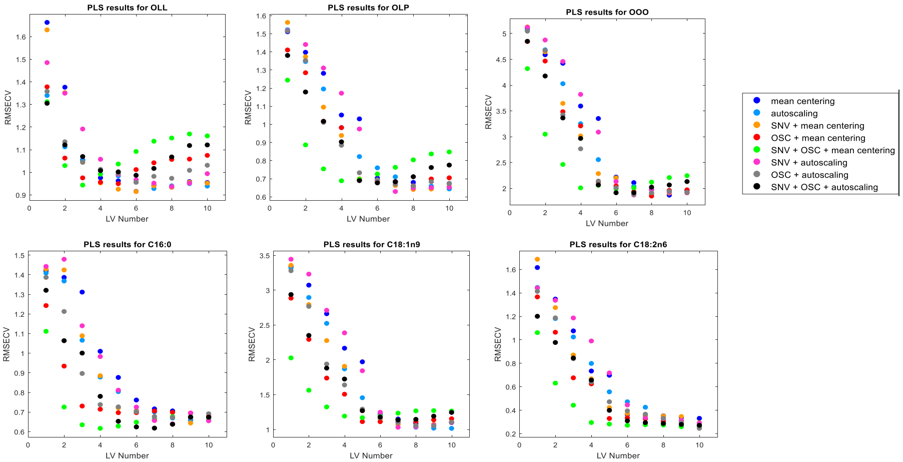

Figure 1.3.1: PLS regression models of NIR spectra acquiring with quartz cuvettes for evaluating eight combinations of data pre-processing

The same considerations can be made when comparing the results obtained by the modelling of DGV spectra: for these models, the combination of SNV + OSC + mean centring (represented in green in figure 1.3.2) 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 1.3.3, original spectral profiles and spectra after

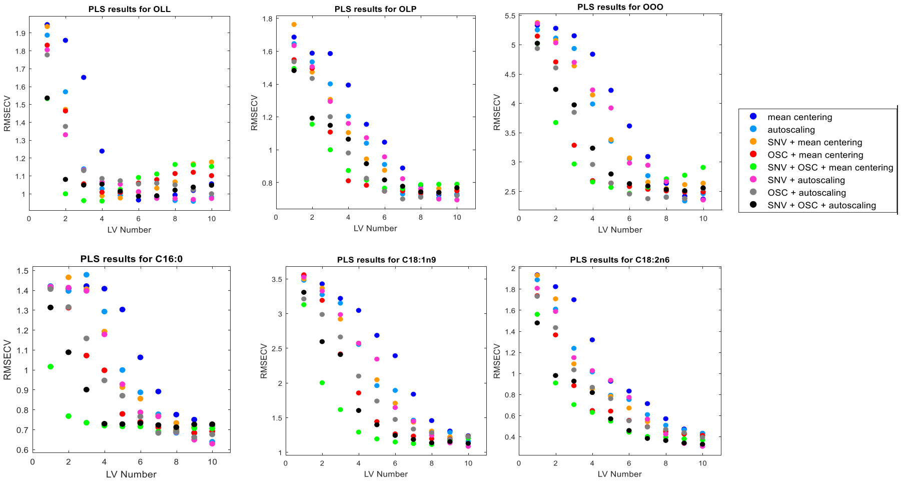

Figure 1.3.2: PLS regression models of NIR spectra acquiring with DGV for evaluating eight combinations of data pre-processing.

pre-treatment, are shown: Figure 1.3.3a shows the raw signals acquired using QC, while Fig. 1.3.3b represents the QC spectral profiles transformed by SNV + OSC for variable C18:1n9. Similarly, Fig. 1.3.3c shows the original signals acquired using GDV and Fig. 1.3.3d the data transformed for the same variable using SNV+OSC. Using two different row pre-treatments as SNV and OSC, it was possible not only to remove the effect caused by interferences of scatter, but also to emphasise the information embodied within the spectra according to the feature that must be modelled. This approach allowed decreasing the number of LVs to retain and therefore the complexity of the models. For a better comparison of raw and transformed profiles, mean centring was not included in this representation.

Table 1.3.1: Calibration and prediction models for quartz cuvettes and disposable glass vials

Quality parameter	Type of cuvettes	Mean	Range (min-max)	Number of LV	RMSECV	RMSECV %	RMSEP	RMSEP %
OOL	QC	13.04	4.29-1.53	4	0.99	7.59	0.75	5.75
	DGV			4	0.96	7.36	0.91	6.98
OLP	QC	6.99	12.92-4.20	4	0.69	9.87	0.68	9.73
	DGV			6	0.76	10.87	1.09	15.56
OOO	QC	38.36	50.22- 23.13	4	2	5.21	1.62	4.22
	DGV			6	2.46	6.41	2.10	5.47
C16:0	QC	12.86	16.40-9.53	4	0.61	4.74	0.58	4.51
	DGV			6	0.71	5.56	0.77	5.97
C18:1n9	QC	72.48	79.32-58.55	5	1.17	1.61	1.2	1.66
	DGV			5	1.19	1.64	1.29	1.79
C18:2n6	QC	7.49	16.38-4.78	5	0.28	3.74	0.28	3.74
	DGV			6	0.44	5.87	0.48	6.41

After choosing the proper data pre-treatment, PLS models were validate on samples belonging to the test set. The model parameters, calculated on pre-processed spectra, are presented in Table 1.3.1 for both QC and DGV data. For each quality parameter a direct comparison between QC and DGV model can be performed in term of number of LVs selected, error in cross-validation and in prediction. In more detail, RMSE are reported in the corresponding variable unit and also as percentage calculated in respect to the mean. The percentage value allows a direct understanding of model goodness.

For some of the models presented the results obtained, in term of predictive capability, cannot be considered completely satisfactory. This is due to the fact that the reduced variability in the EVOO samples for the content of FAMEs and TAGs, did not allow obtaining PLS regression models with good predictive performances. Looking at the results, it was possible to notice that they seem slightly better for models calculated using QC, but a numerical comparison between RMSECV% and RMSEP% of the PLS models is not meaningful to understand if the analytical 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 method was performed on the test set data. The null hypothesis (H0) was that the slope is not significantly different from 1 and that the intercept was not significantly different from 0 at a 95% confidence level; the results of the Passing-Bablok regression are presented in Table 1.3.2. For sake of completeness, for both slope and intercept, limit of acceptability (LL =lower limit and UL= upper limit) and calculated value (CAL) are reported.

Table 1.3.2: Passing-Bablok regression results related to a joint test on slope and intercept values of the regression lines, 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	Accepted
OLP	1.08	1.82	1.40	-5.30	-0.41	-2.53	Accepted
OOO	1.17	1.80	1.48	-30.60	-5.92	-17.83	Accepted
C16:0	0.72	1.17	0.89	-2.18	3.41	1.38	Accepted
C18:1n9	0.81	1.32	0.99	-23.06	14.57	1.42	Accepted
C18:2n6	1.02	2.55	1.55	-11.09	-0.28	-3.97	Accepted

Although QC models seem to better predict EVOO quality parameters, Passing-Bablok test highlighted that there were not statistical differences between models calculated with QC and those obtained with DGV; the null hypothesis (H0) was in fact accepted for all six parameters (OOL, OLP, OOO, C16:0, C18:1n9, C18:2n6) considered. Considering these results, it was possible to state that comparable results were obtained for FAMEs and TAGs prediction with both quartz cuvettes and disposable glass vials.

1.3.4 Conclusions

In order to optimize the timing of NIR acquisition for olive oil routine 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 spectra used to build PLS calibration models for predicting some EVOO quality parameters.

Thanks to a Passing-Bablok test it was possible to highlight that there are not statistical differences between models calculated with QC and those obtained with DGV; this statement was demonstrate for all six the parameters (OOL, OLP, OOO, C16:0, C18:1n9, C18:2n6) considered. Considering these results, the employment of DGV for recording NIR spectra would bring greater benefits for screening analysis of olive oil samples rather than to quantify low concentrations of analyte. In order to understand if DGV can replace QC also for different analysis, this study will be extended, measuring other quality parameters commonly used for routine analysis of extra virgin olive oil such as free acidity or peroxide value.

Acknowledgements

We thank AGER, Project code: 2016-0169 for economic support and Buchi for supplying the mono-use glass cuvettes.

1.3.5 References

1. International Olive Oil Council (2013) Trade standard applying to olive oil and olive pomace oil. In: (RES. COI/T.15/NC no. 3/Revision 7).
2. E. Forina, M., Armanino, C., Lanteri, S. & Tiscornia, Classification of olive oils from their fatty acid composition, Food Research and Data Analysis. (n.d.) 189-214.
3. L. Schwingshackl, G. Hoffmann, Monounsaturated fatty acids , olive oil and health status : a systematic review and meta-analysis of cohort studies, Lipids in Health and Disease. (2014) 154.
4. M. Casale, P. Oliveri, C. Casolino, N. Sinelli, P. Zunin, C. Armanino, M. Forina, S. Lanteri, Analytica Chimica Acta 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 (2012) 56–63. https://doi.org/10.1016/j.aca.2011.11.015.
5. M. Blanco, I. Villarroya, NIR spectroscopy : a rapid-response analytical tool, TrAC Trends in Analytical Chemistry. 21 (2002) 240–250.
6. R.J. Mailer, Rapid Evaluation of Olive Oil Quality by NIR Reflectance Spectroscopy, Journal of the American Oil Chemists' Society. 81 (2004) 823–827.
7. M. Manley, K. Eberle, 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, 126 (2006) 111–126.
8. I. Wesley, E. Pacheco, A.E.J. Mcgill, Identification of Adulterants in Olive Oils, (1996) 515–518.
9. G. Downey, P.M.C. Intyre, A.N. Davies, Geographic Classification of Extra Virgin Olive Oils From the Eastern Mediterranean by Chemometric Analysis of Visible and Near-Infrared Spectroscopic Data, 57 (2003).
10. S. Wold, L.E. Sjostrom, Michael, PLS-regression : a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems. Volume 58 (2001) 109–130.
11. P. Oliveri, C. Malegori, R. Simonetti, M. Casale, Analytica Chimica Acta The impact of signal pre-processing on the final interpretation of analytical outcomes e A tutorial, Analytica Chimica Acta. 1058 (2019) 9–17. https://doi.org/10.1016/j.aca.2018.10.055.

12. R. W. Kennard and L. A. Stone, Computer aided design of experiments, Technometrics. (1969) 137–148.
13. J. Wold, Svante, Henrik Anttia, Fredrik Lindgrenb, Orthogonal signal correction of near-infrared spectra, Chemometrics and Intelligent Laboratory Systems. Volume 44 (1998) 175–185.
14. W.B. H. Passing, A New Biometrical Procedure for Testing the Equality of Measurements from Two Different Analytical Methods, Clinical Chemistry and Laboratory Medicine. 21 (1983) 709–720.

Chapter 2

Chemometric Strategies in Environmental Projects

2.1 Project IV

Different analytical approaches for the biomonitoring of air pollution in Liguria region (northwest Italy) by lichens

Summary

Fast, simple and ‘green’ analytical approaches, based on spectroscopic techniques coupled with chemometrics for the biomonitoring of air pollution in Liguria region (northwest Italy) are presented.

For 2 consecutive years, Lichen thalli of *Pseudevernia furfuracea*, collected far from local sources of air pollution, have been exposed to the air for three months in different areas in the Liguria region. The transplanted thalli, retrieved at the end of the exposure period, have been analyzed by Front-Face Fluorescence Spectroscopy (FFFS), Near Infrared Spectroscopy (FT-NIRS) and moreover measurements of fast chlorophyll fluorescence induction kinetics have been performed. A comparison with the values of environmental pollutants recorded during the exposure period by the Regional Agency for Environmental Protection was made, with the final objective of relating pollutants values in lichens with their atmospheric concentrations.

Chemometric evaluation of the spectra included principal component analysis and quadratic discriminant analysis; the prediction rate of the discriminant models calculated on the emission spectra ranged from 71-80% on external test sets indicting front-face fluorescence spectroscopy as a promising technique for the determination of level and type of pollutants in lichen thalli.

2.1.1 Introduction

Lichens are symbiotic associations between a fungal partner, the mycobiont, and one or more photosynthetic partners, the photobiont, which can be either a green alga or a cyanobacterium [1]. Lacking organs for active water uptake, structures for regulating gas exchanges and permeability barrier for water, lichens are susceptible to absorb water, nutritive substances and gases directly from the atmosphere. Thus, they are extremely sensitive to the presence of substances that alter the atmospheric composition (e.g. SO2 and NOx) and are among the most widely used biomonitors of air pollution [2].

For biomonitoring studies, lichens may be used as bioaccumulators, to estimate the accumulation of trace elements within the lichen thalli over space and time [2], or a bioindicators, to assess any alteration of the community diversity and composition [3] and to estimate changes of physiological biomarkers in response to atmospheric pollutants

[4-5]. From a physiological perspective, it has been widely demonstrated that the exposure of lichens to many gaseous pollutants (i.e. SO2 and NO2) may causes membrane injury, ultrastructural alterations, pigment degradation and/or impairment of photosynthetic function [6-8]Conventionally, these biomarkers may be evaluated by means of spectrophotometric or fluorimetric techniques. Recently, the assessment of the efficiency of the photosynthetic process in the algal population is one of the most common biomarkers used [8-10]. The use of direct light fluorimeter (Plant Efficiency Analyser, PEA) allows obtaining information on the efficiency of the photosynthetic processes on the tylacoid membranes of the algal chloroplasts, from the connectivity between PSII reaction centers to the electron flow to PSI. Particularly, PEA records the maximum

quantum yield of primary photochemistry of the photobiont (measured by the ratio F_V/F_M) and other fluorescence parameters, which can be consider as highly sensitive and reliable tools for studying changes in photosynthetic apparatus and in its working efficiency caused by the negative effects of atmospheric pollution. Differently, when consider lichen as a bioaccumulator, we can obtain information on their trace elements content, thus on the atmospheric contaminants.

The main conventional analytical techniques used to determine element concentration consist of atomic absorption spectrophotometry techniques such as ICP-AES and ICP-MS [11]. Although these techniques are accurate and reliable in giving a quantitative result, they require long laboratory procedures and they are not able to establish unambiguously a relation between any change in the lichen physical and chemical properties and the individuals pollutants in the atmosphere [12].

In this paper, we tested an alternative approach, which combining information from different analytical sources, could potentially provide a comprehensive evaluation of the complex chemical phenomena that occur in complex matrices. For this reason, spectroscopic techniques (e.g. visible (VIS), near infrared (NIR) and mid infrared (MIR) spectroscopy) were considered in order to integrate the assessment of atmospheric pollution by means of lichens. Spectroscopic analysis exploits the interaction of electromagnetic radiation with atoms and molecules to provide qualitative and quantitative chemical and physical (structural) information that is contained within the wavelength or frequency spectrum of energy that is either absorbed or emitted. Spectroscopy in the visible, near and mid-infrared ranges is an increasingly growing technique due to its cost, rapidity, simplicity, and safety, as well as its ability to measure multiple attributes simultaneously without monotonous sample preparation, making it suitable to be implemented on a routine basis. Near infrared spectroscopy (NIRS),

Front-Face Fluorescence Spectroscopy (FFFS) and Plant Efficiency Analyser (PEA) are not expensive and ‘green’ because no reagents are required and thus no waste is produced.

By using the application of mathematical and statistical techniques, chemometrics allows to extract chemical and physical information from complex multidimensional data [13], which are currently observed in spectroscopy techniques. Chemometrics often relies on visualization to help the chemist to obtain the required information, and the most used method in this respect is principal component analysis (PCA). PCA extracts information from data tables by transforming them into plots [14].

In our previous work [15], we showed that NIR spectroscopy coupled with chemometrics was able to generate a lichen ‘fingerprint’ capable of discriminating between samples exposed in a polluted or non-polluted area. Differently, FFFS is usually applied on food samples for classification purposes [16-17], whereas, according to our knowledge, this technique was not investigated for lichen biomonitoring.

The present study aimed at testing the use of different analytical spectroscopic approaches, coupled with chemometrics, as rapid and simple tools for assessing effects of air pollutants on lichen thalli. For achieving this goal, thalli of the fruticose lichen *Pseudevernia furfuracea* (L.) Zopf v. *furfuracea*, collected from a pristine area, have been transplanted for three months to 15 sites in the Liguria region (NW-Italy), characterized by contrasting levels and type of atmospheric pollution, as measured by the regional Environmental Protection Agency (ARPAL). Lichen samples have been analyzed by FFFS, NIRS and PEA and data elaborated by multivariate data analysis (chemometrics), in order to compare the performances of these spectroscopic techniques and to highlight possible synergic or complementary information.

2.1.2 Material and Methods

2.1.2.1. Sample and Reagents

The fruticose epiphytic lichen Pseudevernia furfuracea (L.) Zopf v. *furfuracea* was selected because it is widely used in biomonitoring studies with transplants [18-23].

Lichen thalli were collected from northerly exposed barks of Picea abies (L.) H. Karst in a forest area of Valtournenche (Valle d'Aosta, Italy) at 1900 m a.s.l., far from local sources of air pollution [24]. Collecting lichens from the north side of tree allows work with material adapted to homogeneous regime of diffuse light [25]. Samples were picked up, at 1.5 - 2.0 m above the ground, together with a piece of the supporting branch, using garden shears. The material was taken to the laboratory in paper bags and left to dry out at room temperature and low light overnight (≈5 μmol m^{-2} sec^{-1}), to minimize a rise in the F_V/F_M caused by recovery from natural photoinhibition [26]. Samples were divided into two groups: one, including samples that were never exposed in the experimental sites, were kept in freezer until the end of the experiments (control), whereas the second group included one hundred and fifty thalli which were randomly selected and prepared to be exposed in the 15 exposure sites. In the laboratory, lichen thalli were fixed by means of plastic bands on plastic nets (of ca. 25 × 15 cm) and put into paper bags.

2.1.2.2. Study Area and Sampling Sites

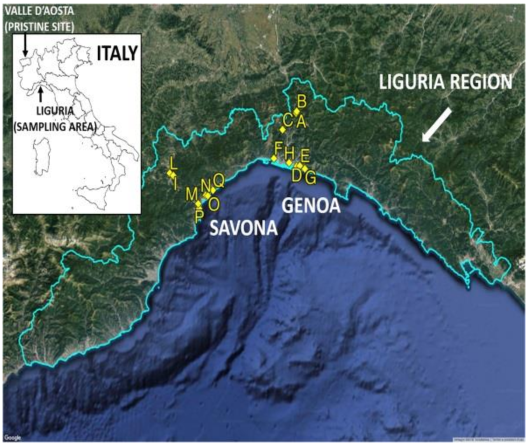

Figure 2.1.1: map of the 15 exposure sites in the Liguria region.

Fifteen sites (A - Q) distributed in an area of ca. 200 Km^2 in Liguria region (NW Italy) (Fig. 2.1.1) were selected for exposure. Particularly, lichen samples were exposed in the urban e industrial area of Genoa (D - H) and Savona (M - Q) and in their hinterland (A - C, and I - L, respectively) (Table 2.1.1). Site A and B, characterized by high levels of air pollution, were located in two hinterland districts of the province of Genoa nearby the highway and an oil refinery. Like the two previous ones, site C was located in the hinterland of Genoa but it differs from the previous one because it is mainly an urban area characterized by a lower level of air pollution. Site D and E were located in the city center of Genoa near the principal traffic congested roads, whereas site F and H were located in the city of Genoa near the industrial harbours and close to the shipyards. Site G was in a little green area in the center of Genoa surrounded by a small traffic road. Site I and L were in the hinterland of Savona (NW Liguria), the first in a small village with low traffic and the other near a big industrial settlement. Finally, site L - Q were in the urbanized area of Savona subjected to different traffic density.

2.1.2.3 Sample Exposure

In each experimental site, 10 lichen thalli (fixed on three plastic nets as described above) were attached on the trunk of adjacent three trees, at approximately 2.5m above the ground, protected by the canopy from direct sunlight.

The sampling was performed for 2 consecutive years, 2015 and 2016, in order to take into account the temporal variability. Fig. 2.1.1 shows the map of the 15 exposure sites in the Liguria region (NW Italy).

In 2015, the lichen deployment was carried out in two consecutive days in July. Lichen samples were transplanted to 15 sites, close to (<50 m) the monitoring stations of the Liguria Regional Environmental Protection Agency (ARPAL, http://www.arpal.gov.it), in the province of Genoa and Savona, according to expected contrasting levels of atmospheric pollution. Thirteen thalli were not exposed and thus considered as control samples. In the second year, the experimental effort was reduced on the basis of the information provided by the results obtained in the first year. Accordingly, in July 2016 only 5 of the 15 stations monitored during 2015 were selected as a representative set, in terms of level and type of atmospheric pollution and geographical location. Only 6 thalli were reserved for controls.

In both years, the sampling lasted three months. Hereafter, samples were retrieved packed in paper bags, protected from sunlight, and transported back to the laboratory, where they were detached from branches, carefully cleaned from debris and dead or senescent parts, and kept in dark conditions at ambient temperature until analysis. Unfortunately, in both sampling years not all thalli were found at the end of the exposure period of three months. Table 2.1.1 shows the list of the remaining samples that were analyzed. For some stations, the number of samples analyzed by FFFS, NIRS and PEA can be different.

2.1.2.4 Air Monitoring Pollution Data

Data on concentrations of the main air pollutants (Benzene, NO_2, SO_2 and PM_{10}) were continuously (hourly) recorded in each experimental site and over the entire exposure periods (2015 and 2016 campaigns) by the devices of the Liguria Regional Environmental Protection Agency (ARPAL), located close to the transplanted thalli. In Table 2.1.1, for each site, we reported the hourly average concentrations of the main air pollutants (Benzene, NO_2, SO_2 and PM_{10} expressed in mg/m3) recorded by the ARPAL during the 3 months of exposure. These data were used to categorize the sampling sites on the basis of their level and type of pollution.

2.1.2.5 Instrumental

All the instrumental measurements described in this section were performed on control (i.e., not exposed) and on transplanted thalli, (i.e., at the end of exposure periods). For the FFFS and NIR analyses, each lichen sample was firstly pulverized with a ball mill, and then the powder was divided in two portions: 0.5 g were used for FFFS analysis and 1.0 g for NIR measurements. Differently, PEA analyses were performed directly on the top of the lacinia of lichen thalli.

2.1.2.5.1 Front-Face Fluorescence Spectroscopy (FFFS)

Emission spectra were recorded using a PerkineElmer LS55 (Perkin-Elmer Ltd., Beaconsfield, U.K.) luminescence spectrometer equipped with a Xenon lamp and a variable angle front-surface accessory. The incidence angle of the excitation radiation was set at 56° to ensure that reflected light

Table 2.1.1: List of the experimental sites (locality) with the corresponding code (site code). For each site are reported summary of samples analyzed in 2015 and 2016 (N analyzed samples), of the average concentrations of the main air pollutants (Benzene, NO2, SO2 and PM10 expressed in mg/m3) recorded by the ARPAL in the 3 months of exposure (pollution data) and relative pollution category.

Site code	Locality	N analysed samples						Pollution data								
		2015			2016			2015				2016				
		FFFS	NIR	PEA	FFFS	NIR	PEA	NO_2	SO_2	PM10	Benzene	NO_2	SO_2	PM10	Benzene	Pollution category
CTR	Controls, not exposed	13	13	13	6	6	6									Control
A	Busalla-Piazza Garibaldi	7	7	7				36.2	9.2		1.38	31.5	10.0	23.8	1.29	Industrial
B	Busalla-Sarissola	7	4	6	10	6	6	15.8	7.2		0.80	26.7	8.3		1.22	Industrial
C	Campomorone	6	4	6							1.59	22.2	6.0	12.9	0.37	Low traffic
D	Genova-Corso Buenos Aires	9	9	8				40.2		28.8	1.71	40.1		31.7	1.80	High traffic
E	Genova-Corso Europa	7	3	7	9	9	5	44.4		35.4	1.70	51.4		29.6	2.19	High traffic
F	Genova-Multedo	4	4	4					9.4		1.20		9.1		0.71	Industrial
G	Genova-Quarto	6	6	6				28.1	3.9	16.0	0.66	11.2	4.7	14.0	0.66	Low traffic
H	Genova-Via Buozzi	6	6	6	9	6	5	48.9	12.3		1.70	50.1	7.9		1.49	Industrial
I	Cairo Montenotte-Loc Bragno	6	6	6				10.0	5.7	17.5	0.90	10.3	5.4	17.9	1.01	Low traffic
L	Cairo Montenotte-Loc Mazzucca	8	6	8	4	4	2	14.6	8.1	19.5	1.62	15.4	9.7	19.3	2.07	Industrial
M	Quiliano	6	3	6				14.6	3.0	21.9	0.79	15.3	4.0	20.9	0.35	Low traffic
N	Savona-Corso Ricci	7	7	6				25.8		12.6	1.28	29.6		16.5	1.33	Low traffic
O	Varaldo	6	3	6				11.0	5.0	17.6	0.37	13.4	3.9	17.6	0.20	Low traffic
P	Vado Ligure	3	3	3	10	6	6	37.5	5.1	24.8	3.32	25.4	4.8	20.7	1.83	High traffic
Q	Albissola	7	7	7						15.2	1.44			15.4	1.32	Low traffic

and scattered radiation were minimized. The incidence angle of the excitation radiation was set at 56° to ensure that reflected light and scattered radiation were minimized. Samples were placed in cuvettes with a circular surface of diameter 15 mm. Excitation and emission slits were both set at 10 nm. Emission spectra were recorded between 300 and 500 nm (with 0.5 nm resolution) at excitation wavelength of 270 nm. Intensities were plotted as a function of the emissionwavelength. For each sample, measurements were done in triplicate to minimize remaining scattering effects and the average signals were used in the multivariate data analysis. The BL Development software (PerkinElmer) was used to register the fluorescent signals.

2.1.2.5.2 Near Infrared Spectroscopy (NIRS)

NIR measurements were carried out using an FT-Near-Infrared Spectrometer, based on a Polarization Interferometer (Buchi NIRFlex N-500), in the 4000-10,000 cm^{-1} range with 4 cm^{-1} resolution. NIR Operator software (Buchi) was used to register the NIR spectra.

For each sample, approximatively 1 g of powder was placed in an optical glass Petri dish and analyzed in reflectance mode. An average of 64 scans was taken for each spectrum. The optical glass dish was washed in warm water, accurately rinsed and dried before carrying out the three replicates of each sample. The average signals were used in the multivariate data analysis.

2.1.2.5.3 Plant Efficiency Analyser (PEA)

Chlorophyll a fluorescence (Chl a) measurements were performedusing Handy-PEA chlorophyll fluorometer (Plant Efficiency Analyser, Hansatech instruments Ltd, Norfolk, England). Prior to taking the measurements, samples were sprayed with deionized water until wet and adapted to darkness for 15 min. Three Chl a measurements were performed on each thallus. The Chl a fluorescence transients were induced by a red light (peak at 650 nm) provided by an array of three high-intensity LEDs. Data were recorded after a saturating light pulse (3500 μmolnm^{-2} sec^{-1}) of 1 s. The gain of the PEA was 0.8. The fluorescence transient rises from F_0 (when all PSII reaction centers are open, i.e. when the primary acceptor quinone is full oxidized) to F_M (when all the PSII reaction centers are closed, i.e. the full reduction of the primary acceptor quinone). The potential quantum yield of primary photochemistry (F_V/F_M) was calculated as $(F_M-F_0)/F_M$.

Table 2.1.2: Definitions of the OJIP parameters based on Stirbet and Govindjee (2011).

OJIP Parameters	**Description**
F_0	First reliable fluorescence value after the onset of actinic illumination; used as initial value of the fluorescence
F_M	Maximal fluorescence
F_V	Maximum variable Chl fluorescence
F_V/F_M = TR/ABS	Maximum quantum yield of primary PSII photochemistry
T_{fm}	Time to reach the maximum fluorescence value F_M
Area	Area between OJIP curve and the line $F = F_M$
ABS/RC	Average absorbed photon flux per PSII reaction center (or also, apparent antenna size of an active PSII)
TR/RC	Maximum trapped exciton flux per PSII
DI/RC	Energy flux which is dissipated chiefly as heat
ET/RC	Electron transport flux from Q_A to Q_B per PSII
RC/ABS	Number of Q_A reducing RCs per PSII antenna Chl
ABS/CS	Absorbed photon flux per cross section (or also, apparent PSII antenna size)
RC/CS	Number of active PSII RCs per cross section
TR/CS	Maximum trapped exciton flux per cross section
ET/CS	Electron transport flux from Q_A to Q_B per cross section
DI/CS	Heat dissipation per cross section
ET/TR	Efficiency/probability with which a PSII trapped electron is transferred from Q_A to Q_B
RE/ET	Efficiency/probability with which an electron from Q_B is transferred until PSI acceptors
PI	Global indicator used to express the overall vitality of the samples

We also considered other parameters to describe the ability of the photobiont in transferring trapped photons along the tilacoid membrane from PSII to PSI [27-29]. For a detailed description of the parameters and formulae, see Table 2.1.2.

2.1.2.6 Data Analysis

2.1.2.6.1. Data Matrices Organization

Two data matrices were elaborated for each analytical techniques (FFFS, NIRS and PEA): F1, N1 and P1 containing the data relative to year 2015 and F2, N2 and P2 the data of year 2016, respectively.

Regarding FFFS, F1 had 108 rows (samples) and 397 columns (variables acquired between 300 and 500 nm, with 0.5 nm resolution). F2 had the same number of variables as F1 but only 48 rows (samples). F1 and F2 were pre-processed using standard normal variate (SNV) for correcting for shift.

As far as the NIR data are concerned, the part of the spectra from 8000 to 10,000 cm^{-1} was removed since it was not informative, thus N1 had 91 rows (samples) and 1001 columns (variables acquired between 8000 and 4000 cm^{-1}, with 4 cm^{-1} resolution); N2 data matrix had 37 rows (samples) and 1001 columns (variables,between 8000 and 4000 cm^{-1}). NIR spectra were pre-processed using Standard Normal Variate (SNV) to eliminate the unwanted variation due to light scattering. With regards to PEA, three replicates for each thallus were acquired, so that P1 had 315 rows (105 *3) and 21 columns (parameters of efficiency) and P2 had 90 rows (30 *3) and 21 columns.

2.1.2.6.2. Chemometrics Analysis

Principal Component Analysis (PCA) was applied as a data display method on the six spectroscopic data matrices (F1, N1, P1, F2, N2 and P2) and on the pollution data matrix. Quadratic Discriminant Analysis (QDA) was performed as a classification technique on the six spectroscopic data matrices.

PCA is the most used tool in exploratory data analysis and uses an orthogonal transformation to convert a set of correlated variables into a set of uncorrelated variables called principal components [30]. QDA is a probabilistic parametric classification technique, which represents an evolution of Linear Discriminant Analysis (LDA) [31] for nonlinear class separations. Also QDA, like LDA, is based on the hypothesis that the probability density distributions are multivariate normal but, in this case, the dispersion is not the same for all of the categories. It follows that the categories differ not only for the position of their centroid but also for the variance-covariance matrix (different location and dispersion).

For the year 2015, the QDA discrimination rules were validated using both a cross-validation procedure with five cancellation groups (5CV) and an external test set. The test set samples were selected randomly assigning 25% of the samples to the external test and 75% to the training set. For the year 2016, the QDA discrimination rules were validated only in cross validation (5CV) considering the low number of samples. QDA results were expressed as the total prediction rates, that is the ratio of correct predictions to the total number of predictions and it measures the predictive ability.

2.1.3 Results and Discussion

2.1.3.1 Principal Component Analysis

2.1.3.1.1 Air Monitoring Pollution Data

Data collected by ARPAL in the 15 monitoring stations in the province of Genoa and Savona, in the period July-September 2015 are reported in Table

2.1.1. In the study area, the largest sources of SO_2 in the atmosphere include industrial processes, ships and other vehicles emissions and heavy equipment that burn fuel with a high sulfur content [32-33]. Benzene, particles less than 10 mm in diameter (PM_{10}) and nitrogen oxides (NO_x) are the main urban air pollutants due to traffic. Fig. 2.1.2 shows respectively the PCA loading (Fig. 2.1.2a) and score (Fig. 2.1.2b) plots of these pollution data. SO_2 was the only variable showing negative loadings on PC1; NO_2, Benzene and PM_{10}, had positive loadings on PC1. Therefore, PC1 was associated with the type of pollution, 'industrial pollution' at negative values of PC1 and 'pollution from traffic' at positive values. Stations A, B and C, located in the northern hinterland of Genoa, were characterized by a high content of SO_2 and this is potentially due to a soap factory (C) and an oil-refinery producing 700 t/year of SO_2 (A and B). Station H also showed a very high content of SO_2 and this can be explained because this station was in front of the Genoa harbor where many ships dock. Stations E and D are in the most traffic congested streets in the center of Genoa and station P is a touristic area close to the sea and therefore very popular during summer.

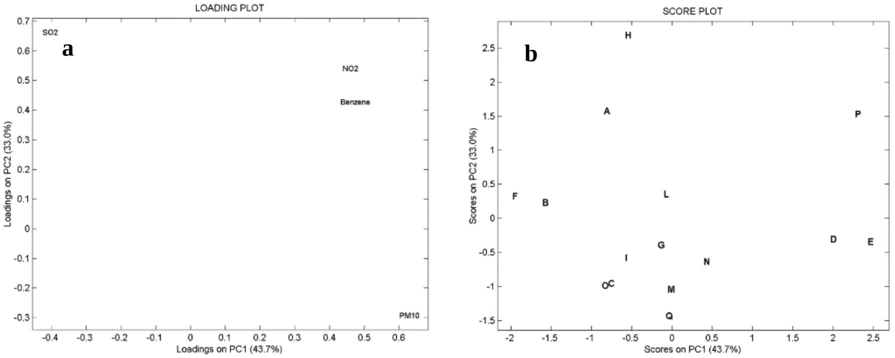

Figure 2.1.2: Loadings (**a**) and score plot (**b**) of pollution data collected by ARPAL in the 15 monitoring stations in the province of Genoa and Savona, in the year 2015.

Fig. 2.1.3 (a and b) shows the PCA loading (S2a) and score (S2b) plots of pollution data collected by ARPAL in the period July-September 2016. The information provided was very similar to that extracted from the 2015 data.

According to the information obtained from the pollution data, samples were divided into 4 classes, in terms of type and level of pollution, which characterized the sites of exposure:

1. Not exposed: Control samples (CTR).
2. Exposed in stations characterized by industrial pollution: A, B, F, H, L
3. Exposed in stations characterized by high-congested traffic: D, E, P
4. Exposed in stations characterized by low traffic: C, G, I, M, N, O, Q.

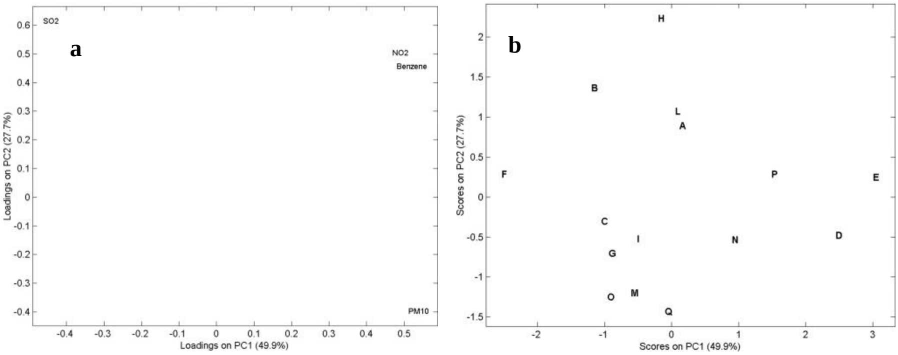

Fig. 2.1.3: Loadings (**a**) and score plot (**b**) of pollution data collected by ARPAL in the 15 monitoring stations in the province of Genoa and Savona, in the year 2016.

Based on the results of the previous survey, for the year 2016 the category “low traffic” was excluded from the analysis.

2.1.3.1.2 FFFS

Fig. 2.1.4 shows the score plot of the data set F1 (year 2015), in the space of the 2 first PCs. The control samples were at higher positive scores on PC1 that explains the 48% of the total variance and they were all clustered, highlighting a good homogeneity of the starting samples. On the contrary, samples exposed in 4 industrial sites (A, B, F and L) were associated with negative scores of PC1. Samples from traffic sites (both high-congested and low traffic) showed a less uniform pattern with respect to samples exposed in industrial sites, however half of the samples from traffic sites (E, M, N, O and P) were associated with negative scores of PC2.

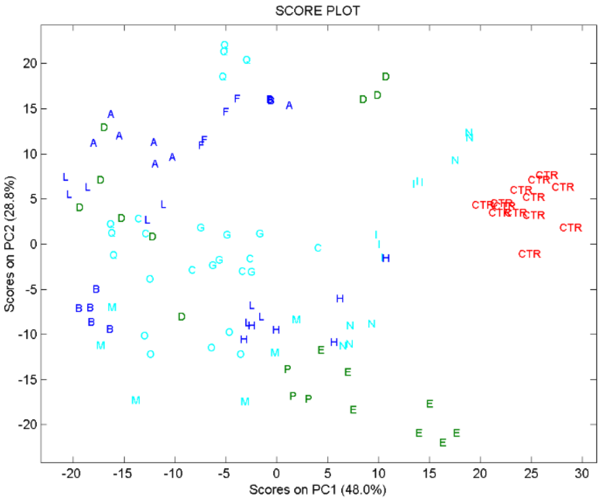

Figure 2.1.4: Score plot of the FFFS emission spectra acquired on the lichens thalli exposed during 2015. Samples are indicated by their pollution classes: red control; blue industrial pollution; green high congested traffic; light blue low traffic.

Fig. 2.1.5 shows the score plot of the data set F2 (year 2016), in the space of the 2 first PCs. Control samples were associated with negative scores of PC2 and form a defined cluster with respect to samples transplanted in the exposed sites. The separation of the pollution classes was evident along PC1: industrial sites were associated to negative scores whereas traffic sites with positive ones.

2.1.3.1.3 NIRS

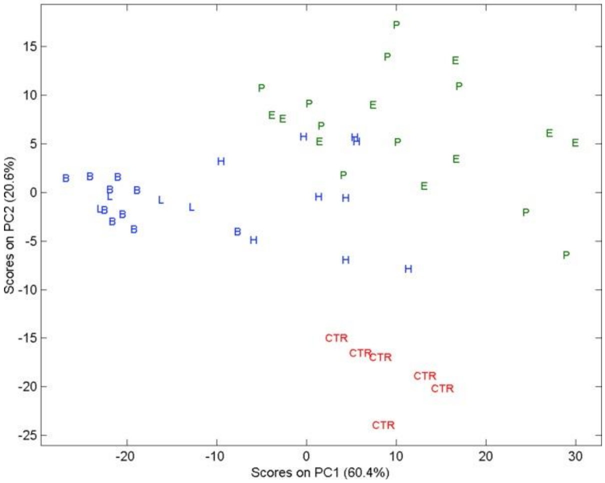

Figure 2.1.5: Score plot of the FFFS emission spectra acquired on the lichens thalli exposed during 2016. Samples are indicated by their pollution classes: red control; blue industrial pollution; green high congested traffic.

Fig. 2.1.6 shows the score plot of the data set N1 (year 2015), in the space of PC1-PC4. Control samples formed a well-defined cluster, associated with

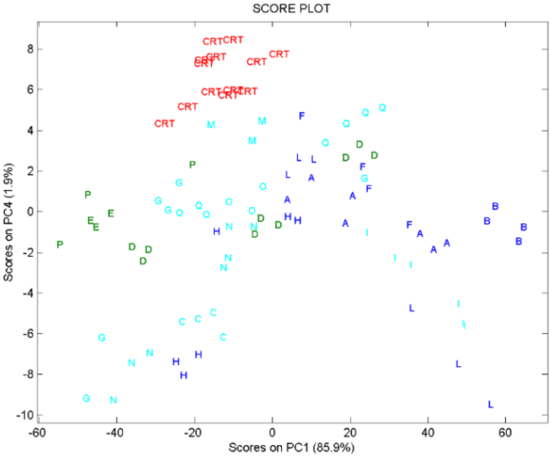

Figure 2.1.6: Score plot of the NIR spectra acquired on the lichens thalli exposed in the year 2015. Samples are indicated by their pollution classes: red control; blue industrial pollution; green high congested traffic; light blue low traffic.

negative scores of PC1 and positive of PC4. Differently, the separation of the pollution classes was less evident. Overall, samples from industrial sites were associated with positive scores of PC1, whereas samples from traffic sites (high congested traffic D, E and P; low traffic C, G, M, N and O) occurred for negative values of PC1.

Fig. 2.1.7 shows the score plot of the data set N2 (year 2016), in the space of PC1-PC2. Control samples were associated with positive scores of PC2. With regard to samples transplanted in the exposed sites we can observe that samples from traffic sites occurred for positive scores of PC1 whereas samples from industrial sites to negative ones.

2.1.3.1.4. PEA

The first 2 PCs of P1 data set explained the 75.1% of the variance (Fig. 2.1.8b). Control samples were associated with negative scores of PC1, corresponding to high photosynthetic efficiency. Samples exposed in 4 low

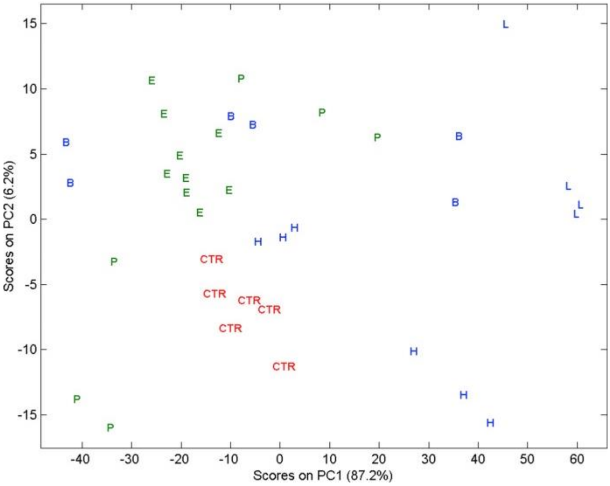

Figure 2.1.7: Score plot of the NIR spectra acquired on the lichens thalli exposed in the year 2016. Samples are indicated by their pollution classes: red control; blue industrial pollution; green high congested traffic.

traffic sites (M, N, O, Q) and in 2 high traffic sites (P, E) were found for positive scores of PC1, associated to low photosynthetic efficiency (Fig. 2.1.8a). The separation of pollution categories along PC2 was much less evident; this component was associated to higher heat dissipation (DI/RC), as a response of lichens to high level of stress [8].

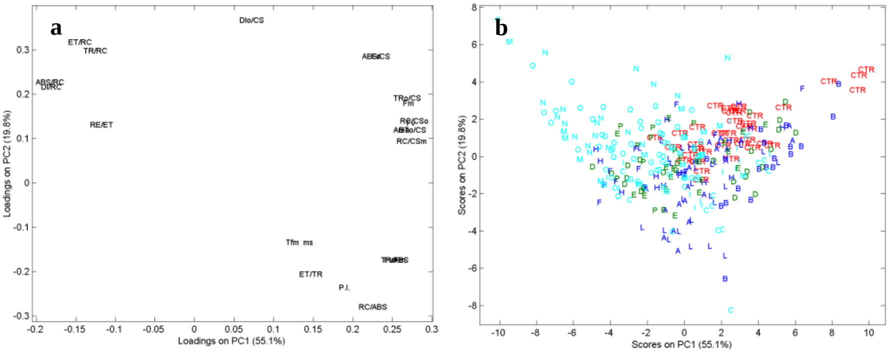

Figure 2.1.8: loadings (**a**) and score plot (**b**) of the PEA values measured on the lichens thalli exposed in the year 2015. Samples are indicated by their pollution classes: red control; blue industrial pollution; green high congested traffic; light blue low traffic.

Figs. 2.1.9 a and b show respectively the loading and score plots for P2 data set. Overall, the results obtained in 2016 were comparable to those of the previous year. PC1 (54.7% of total variation) was associated with an increasing gradient of pollution, ranging from control samples to those transplanted in industrial and high congested traffic sites. These latter samples showed a low photosynthetic efficiency and high heat dissipation, as a response to stressing conditions.

2.1.3.2. Classification

QDA was applied as a classification method on the FFFS, NIR and PEA data of the year 2015 and 2016 (Table 2.1.2), in order to evaluate the possibility to discriminate the lichen thalli categorized on the basis of the level and type of air pollutants (see 2.1.3.1.1). For year 2015, the mean prediction rate of the discriminant rule calculated on the FFFS emission spectra was 70% on the external test set, supporting FFFS as a promising technique for discriminating the effects of different levels and type of pollutants on lichen thalli; in fact, considering the biological variability of the lichen thalli, the QDA results can be considered more than satisfactory.

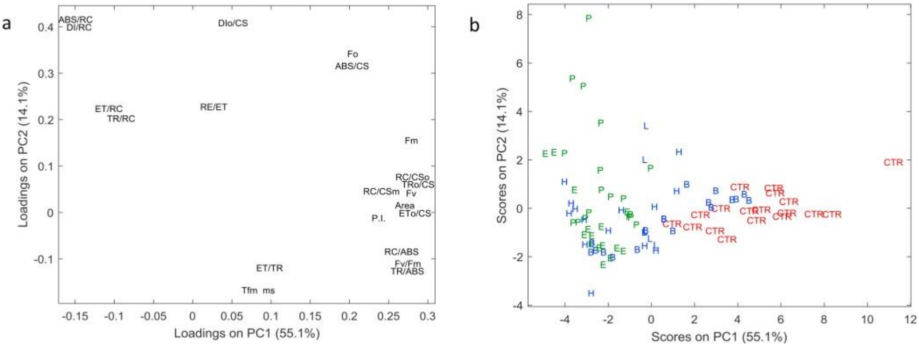

Figure 2.1.9: loadings (**a**) and score plot (**b**) of the PEA values measured on the lichens thalli exposed in the year 2016. Samples are indicated by their pollution classes: red control; blue industrial pollution; green high congested traffic.

On the contrary, the results obtained with the NIR spectra were not so satisfactory. In particular, high prediction rates were obtained only in the discrimination between controls and other classes. This indicates NIR spectroscopy as an analytical method able to differentiate between samples exposed or not in polluted areas [15], but not as sensitive to discriminate between areas with different types of atmospheric pollution.

Table 2.1.2: QDA results on the FFFS (F1 for 2015 and F2 for 2016), FT-NIR (N1 for 2015 and F2 for 2016) and PEA (P1 for 2015 and P2 for 2016) data of the year 2015 and 2016. Results are the mean correct prediction rate expressed as percentages (%).

Data set	Control	Industrial	High-congested traffic	Low traffic	Weighted mean
F1	100.0	80.0	100.0	25.0	70.0
F2	100.0	75.0	71.4	//	75.0
N1	90.0	64.7	66.7	25.0	51.3
N2	83.3	62.5	66.6	//	67.6
P1	100.0	26.1	33.3	45.4	40.0
P2	100.0	66.7	80.0	//	77.8

PEA showed contrasting results, with a low mean prediction rate in 2015 (P1) and a higher performance in 2016 (P2). In both years, PEA successfully discriminated control vs. exposed samples, but it was less efficient in distinguishing between categories of pollution.

As a general result, FFFS showed the highest mean prediction rate and it was able to correctly discriminate sites characterized by different type of pollution (namely, industrial vs. traffic). All techniques showed very high prediction rates for control samples. Moreover, FFFS results were sufficiently reproducible in both years, whereas the other techniques showed discordant results in the two consecutive campaigns. Particularly, PEA was not able to distinguish between polluted categories, even though its performance improved in 2016, when the low traffic category was not taken into account. These outcomes might be due to differences of climatic conditions during the exposure periods, which may have affected the photosynthetic efficiency of transplanted lichens, independently from the atmospheric pollution to which they were subjected. In fact, a synergic effect of climate and pollution may cause a deep alteration of the photosynthetic process. This is in accordance with what was observed by Malaspina et al. [7] who showed that transplanted lichens were highly sensitive to the interaction of atmospheric pollution and proximity to the sea.

Although both FFFS and PEA data are based on fluorescence values recorded on the same samples, they provide different information: PEA data are collected from the intact living organism, while FFFS fluorescence data are based on the chemical properties of the fluorescent compounds in its composition. The results obtained by FFFS and NIRS derive both from myco- and photobiont. From one side, this response is much less specific if compared with the one of PEA, but on the other hand it seems to be less influenced by short-term environmental variations. In lichens, many organic compounds can produce NIR and FFFS absorptions, including e.g., polycyclic aromatic hydrocarbons and organic acids that come from environmental pollution or endogenous organic acids that may increase as a response to stresses such as [15].

2.1.4. Conclusion

Lichen biomonitoring is widely used for detecting air pollution patterns and can be especially useful in remote areas where the use of instrumental recording is hindered by difficult access to sites and difficult management of mechanical and electrical devices.

In this study, the combined use of several rapid analytical approaches, coupled with chemometrics, as rapid and simply tools for assessing the effects of air pollutants on lichen thalli was investigated. Lichen samples were analyzed by FFFS, NIRS and PEA, in order to compare the performances of these analytical spectroscopic techniques, and to highlight possible synergic or complementary information.

Despite the fact that it seems hard to discriminate between similar levels of atmospheric pollution, the explored techniques and in particular FFFS were able to highlight different type of pollution (namely, industrial vs. traffic). Considering the biological variability of the lichen thalli, the classification performances achieved by QDA can be considered more than satisfactory.

This could pose the basis for promising development of spectroscopic techniques for exploring possible range of impact of different sources of emissions in a complex context.

2.1.5 References

1. Nash III, T.H., 2006. Introduction. In: Nash III, T.H. (Ed.), Lichen Biology. Cambridge University Press, pp. 1-8.
2. Bargagli, R., Mikhailova, I., 2002. Accumulation of inorganic contaminants. In: Nimis, P.L., Scheidegger, C., Wolseley, P.A. (Eds.), Monitoring with Lichens - Monitoring Lichens. Kluver Academic Publisher, pp. 65-84. Printed in Netherlands.
3. Giordani, P., 2007. Is the diversity of epiphytic lichens a reliable indicator of air pollution? a case study from Italy. Environ. Pollut. 146, 317-323.
4. Jensen, M., Kricke, R., 2002. Chlorophyll fluorescence measurements in the field: assessment of the vitality of large numbers of lichen thalli. In: Nimis, P.L., Scheidegger, C., Wolseley, W.A. (Eds.), Monitoring with Lichens - Monitoring Lichens. Kluver Academic Publisher, pp. 327-332. Printed in Netherlands.
5. Mikhailova, I., 2002. Transplanted lichens for bioaccumulation studies. In: Nimis, P.L., Scheidegger, C., Wolseley, W.A. (Eds.), Monitoring with Lichens - Monitoring Lichens. Kluver Academic Publisher, pp. 301-304. Printed in Netherlands.
6. Deltoro, V.I., Gimeno, C., Calatayud, A., Barreno, E., 1999. Effects of SO2 fumigations and photosynthetic CO2 gas exchange, chlorophyll a fluorescence emission and antioxidant enzymes in the lichen Evernia prunastri and Ramalina farinacea. Physiol. Plantarum 105, 648-654.
7. Malaspina, P., Giordani, P., Pastorino, G., Modenesi, P., Mariotti, M.G., 2015. Interaction of sea salt and atmospheric pollution alters the OJIP fluorescence transient in the lichen Pseudevernia furfuracea (L.) Zopf. Ecol. Indicat. 50, 251-257.
8. Malaspina, P., Modenesi, P., Giordani, P., 2018. Physiological response of two varieties of the lichen Pseudevernia furfuracea to atmospheric pollution. Ecol. Indicat. 86, 27-34.
9. Calatayud, A., Sanz, M.J., Calvo, E., Barreno, E., Valle-Tascon, S., 1996. Chlorophyll a fluorescence and chlorophyll content in Parmelia quercina thalli from a polluted region of northern Castellon (Spain). Lichenol. 28, 49-65.
10. Maxwell, K., Johnson, G.N., 2000. Chlorophyll fluorescence - a practical guide. J. Exp. Bot. 51, 659-668.
11. Bargagli, R., Nimis, P.L., 2002. Guidelines for the use of epiphytic lichens as biomonitors of atmospheric deposition of trace elements. In: Nimis, P.L., Scheidegger, C., Wolseley, P.A. (Eds.), Monitoring with Lichens – Monitoring Lichens. Kluver Academic Publisher, pp. 295e299. Printed in Netherlands.
12. Nimis, P.L., Scheidegger, C., Wolseley, P.A., 2002. Monitoring with Lichens - Monitoring Lichens. Kluver Academic Publisher (Printed in Netherlands).

13. Wold, S., Sjöström, M., 1972. Statistical analysis of the Hammett equation: 1. Methods and model calculations. Chem. Scripta 2, 49-55.
14. Massart, D.L., Vander Heyden, Y., 2004. From tables to visuals: principal component analysis, Part 1. LC-GC Eur. 17 (11), 586-591.
15. Casale, M., Bagnasco, L., Giordani, P., Mariotti, M.G., Malaspina, P., 2015. NIR spectroscopy as a tool for discriminating between lichens exposed to air pollution. Chemosphere 134, 355-360.
16. Ruoff, K., Luginbühl,W., Künzli, R., Bogdanov, S., Bosset, J.O., von der Ohe, K., von der Ohe, W., Amad_o, R., 2006. Authentication of the botanical and geographical origin of honey by front-face fluorescence spectroscopy. J. Agric. Food Chem. 54, 6858-6866.
17. S_adeck_a, J., T_othov_a, J., M_ajek, P., 2009. Classification of brandies and wine distillates using front face fluorescence spectroscopy. Food Chem. 117, 491-498.
18. Blasco, M., Dome~no, C., L_opez, P., Nerín, C., 2011. Behaviour of different lichen species as biomonitors of air pollution by PAHs in natural ecosystems. J. Environ. Monit. 13, 2588-2596.
19. Kodnik, D., Carniel, F.C., Licen, S., Tolloi, A., Barbieri, P., Tretiach, M., 2015. Seasonal variations of PAHs content and distribution patterns in a mixed land use area: a case study in NE Italy with the transplanted lichen Pseudevernia furfuracea. Atmos. Environ. 113, 255e263.
20. Nascimbene, J., Tretiach, M., Corana, F., Lo Schiavo, F., Kodnik, D., Dainese, M., Mannucci, B., 2014. Patterns of traffic polycyclic aromatic hydrocarbon pollution in mountain areas can be revealed by-lichen biomonitoring: a case study in the Dolomites (Eastern Italian Alps). Sci. Total Environ. 475, 90-96.
21. Sorbo, S., Aprile, G., Strumia, S., Castaldo Cobianchi, R., Leone, A., Basile, A., 2008. Trace element accumulation in Pseudevernia furfuracea (L.) Zopf exposed in Italy's so called Triangle of Death. Sci. Total Environ. 407, 647-654.
22. Tretiach, M., Adamo, P., Bargagli, R., Baruffo, L., Carletti, L., Crisafulli, P., Giordano, S., Modenesi, P., Orlando, S., Pittao, E., 2007a. Lichen and moss bags as monitoring devices in urban areas. Part I: influence of exposure on sample vitality. Environ. Pollut. 146, 380-391.
23. Tretiach, M., Candotto Carniel, F., Loppi, S., Carniel, A., Bortolussi, A., Mazzilis, D., Del Bianco, C., 2011. Lichen transplants as suitable tool to identify mercury pollution from waste incinerators: a case study from NE Italy. Environ. Monit. Assess. 175, 589-600.
24. Malaspina, P., Giordani, P., Modenesi, P., Abelmoschi, M.L., Magi, E., Soggia, F., 2014. Bioaccumulation capacity of two chemical varieties of the lichen Pseudevernia furfuracea. Ecol. Indicat. 45, 605-610.

25. Tretiach, M., Piccotto, M., Baruffo, L., 2007b. Effects of ambient NOx on chlorophyll a fluorescence in transplanted Flavoparmelia caperata (lichen). Environ. Sci. Technol. 41, 2978-2984.
26. Gauslaa, Y., Solhaug, K.A., 2004. Photoinhibition in lichens depends on cortical characteristics and hydration. Lichenol. 36, 133-143.
27. Strasser, R.J., Srivastava, A., Tsimilli-Michael, M., 2000. The fluorescence transient as a tool to characterise and screen photosynthetic samples. In: Yunus, M., Pathre, U., Mohanty, P. (Eds.), Probing Photosynthesis: Mechanisms, Regulation and Adaptation. Taylor & Francis, London, pp. 445-483.
28. Strasser, R.J., Tsimilli-Michael, M., Srivastava, A., 2004. Analysis of the fluorescence transient. In: Papageorgiou, G.C., Govindjee (Eds.), Chlorophyll Fluorescence: a Signature of Photosynthesis, Advances in Photosynthesis and Respiration Series (Govindjee, Series Ed.). Springer, Dordrecht, pp. 321-362.
29. Tsimilli-Michael, M., Eggenberg, P., Biro, B., Köves-Pechy, K., Vörös, I., Strasser, R.J., 2000. Synergistic and antagonistic effects of arbuscular mycorrhizal fungi and Azospirillum and Rhizobium nitrogen-fixers on the photosynthetic activity of alfalfa, probed by the polyphasic chlorophyll a fluorescence transient O-J-I-P. Appl. Soil Ecol. 15, 169-182.
30. Wold, S., 1987. Principal component analysis. Chemometr. Intell. Lab. Syst. 2, 37-52.
31. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers- Verbeke, J., 1998. Supervised pattern recognition. In: Handbook of Chemometrics and Qualimetrics: Part B. Elsevier, Amsterdam, pp. 207-241.
32. Mazzei, F., D'Alessandro, A., Lucarelli, F., Nava, S., Prati, P., Valli, G., Vecchi, R., 2008. Characterization of particulate matter sources in an urban environment. Sci. Total Environ. 401, 81-89.
33. Schembari, C., Cavalli, F., Cuccia, E., Hjorth, J., Calzolai, G., Pérez, N., Pey, J., Prati, P., Raes, F., 2012. Impact of a European directive on ship emissions on air quality in Mediterranean harbours. Atmos. Environ. 61, 661-669.

2.2 Project V:

The Effect of Extraction Methodology on the Recovery and Distribution of Naphthenic Acids of Oilfield Produced Water

Summary

Comprehensive chemical characterization of naphthenic acids (NAs) in oilfield produced water is a challenging task due to sample complexity. The recovery of NAs from produced water, and the corresponding distribution of detectable NAs are strongly influenced by sample extraction methodologies. In this study, we evaluated the effect of the extraction method on chemical space (i.e. the total number of chemicals present in a sample), relative recovery, and the distribution of NAs in a produced water sample. Three generic and pre-established extraction methods (i.e. liquid-liquid extraction (Lq), and solid phase extraction using HLB cartridges (HLB), and the combination of ENV+ and C8 (ENV) cartridges) were employed for our evaluation. The ENV method produced the largest number of detected NAs (134 out of 181) whereas the HLB and Lq methods produced 108 and 91 positive detections, respectively, in the tested produced water sample. For the relative recoveries, the ENV performed better than the other two methods. The uni-variate and multi-variate statistical analysis of our results indicated that the ENV and Lq methods explained most of the variance observed in our data. When looking at the distribution of NAs in our sample the ENV method appeared to provide a more complete picture of the chemical diversity of NAs in that sample. Finally, the results are further discussed.

2.2.1 Introduction

Naphthenic acids (NAs) are naturally occurring compounds in petroleum, with a highly variable composition depending on the source of the oil [1]. The concentration of NAs in petroleum can range from non-detectable to 3% by weight.2 NAs constitute a complex mixture of chemicals, due to the multiple possible chemical structures (i.e. structural isomers) for the same chemical formula. For example for an NA with the formula of $C_{10}H_{18}O_2$, assuming 6 component rings, there are more than 37 isomers. Many of these isomers have a similar structure and thus similar chemical and physical properties. Therefore, a mixture of NAs becomes an extremely challenging matrix to resolve and characterize [2]. As a consequence, the composition of NAs in a complex matrix such as oilfield produced water (PW) is unknown.

Oil production PW is one of the largest streams of industrial treated wastewater in the world [3]. PW is an unresolved complex mixture and consists of a wide variety of chemicals from metals to organic pollutants, including NAs [3–7]. Moreover, multiple studies have reported that the NAs are one of the toxic components of the oilfield PW to a variety of organisms [2,3,8–10]. For example, NAs have been shown to be weak estrogen receptor agonists and androgen receptor antagonists [3,10–12]. Little is, however, known about the chemical composition NAs as well as their environmental fate and behaviour. Consequently, an effective assessment of the risk they pose to the environments receiving oilfield PW difficult. An understanding of the chemical composition of the NAs in the oilfield PW is therefore warranted. The chemical characterization of NAs in PWs is typically performed on the acidic fraction of the total extract of PW [2–4,9]. Typically, liquid-liquid extraction, solid phase extraction, or a combination of both is used in order to tackle the sample complexity provided by both the NAs and PW [2,13-14]. The extraction method used to produce these extracts are compared/ validated either via total extractable material measurement or through the use of a limited number of surrogates as reviewed by Kovalchik et al [13,15–17]. Both mentioned methods have shown to be unable to comprehensively assess the extraction efficiency of one method compared to another [2, 13]. For example, in our previous study we demonstrated that the choice of the extraction procedure changes the explored chemical space of the sample [18]. In that study even though two out of three extraction methods showed similar performance for the surrogate chemicals, more detailed chemical characterization revealed substantial differences among tested extraction methods. However, that study was focused on the volatile and semi-volatile fraction of PW. With regards to NAs, to our knowledge there has not been a detailed extraction recovery assessment based on individual NAs.

To answer that question, we employed three generic and well established extraction methods a liquid-liquid extraction method and two solid phase extraction (SPE) approaches to assess the relative recoveries each NA. We evaluated the effect of each extraction method on both the distribution and the relative recoveries of NAs in PW. The extracts were analysed as such (i.e. no fractionation) via liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS), which was essential to accurate identification of NAs in the PW samples [19].

2.2.2 Material and Methods

2.2.2.1 Sample and the Reagents

A sample of PW (total volume of 5 L) was obtained from an oil platform in the Halten bank off coast of mid-Norway in February 2017 [20]. The sample was divided into 9 aliquots, each of 400 mL. These samples were extracted using three generic extraction methods: liquid-liquid extraction (Lq); Hydrophilic-Lipophilic-Balanced cartridges, here referred to as HLB; and the combination of C8 and ENV+ cartridges, which we refer to as ENV. The HLB cartridges were a combination of two monomers, the hydrophilic N-vinylpyrrolidone and the lipophilic divinylbenzene whereas the ENV cartridges consisted of hydroxylated polystyrene-divinyl benzene copolymer. Both of these methods are considered wide range extraction methods for a combination of polar and non-polar chemicals. The details of the extraction procedure for all three methods are provided elsewhere [18]. In short, the Lq method was the dichloromethane (DCM) extract of the acidified PW, repeated three times, with a final volume of 2 mL. A solution of 1N hydrochloric acid was used for acidification of the PW samples. For the solid phase extraction methods (SPE), both cartridges were conditioned with a combination of methanol and water as recommended by the vendors. The preconditioned cartridges then were loaded with 400 mL of PW using a vacuum pump. These, then, were eluted with two times the volume of the cartridges employing a mixture of hexane, DCM, and 2-propanol. This mixture was selected based on the fact that it appeared inert towards the extracted NAs. The final extracts of 2 mL were stored in the freezer until the analysis. This combination of eluents was previously shown to be effective for extraction of analytes with a wide range of chemical and physical properties in complex samples [18].

Three procedural blanks were generated for each extraction method. For Lq method, these blanks were the extract of the glassware using a mixture of DCM and a 1N solution of HCl. Regarding the SPE methods, the blanks were the extracts of the preconditioned cartridges with the same solvent mixture used for extraction of the samples.

The final extracts, including the blanks, were spiked with 100 ng of diazepam-D5 as the injection standard for monitoring the instrument performance during the analysis.

ACS grade methanol, 2-propanol, hexane, dichloromethane, NH_4OH, and diazepam-D5 were obtained from Sigma-Aldrich, Norway. HPLC grade water was purchased from Waters (Mil- ford, MA, USA).

We obtained the Oasis HLB 6 mL Cartridges, with 200 mg of sorbent from Waters, Norway whereas the ENV+ cartridges, having 100 mg of the sorbent and a total volume of 6 mL, were purchased from Biotage, Sweden. Finally, the C8 sorbent came from Sigma- Aldrich, Norway.

2.2.2.2 Instrumental

Seven µL of each extract was injected into a Waters Acquity UPLC system (Waters Milford, MA, USA) equipped with UPLC HSS C18 column (2.1150 mm, particle size 1.8 mm) (Waters, Milford, MA, USA). The extracts were separated using the following chromatographic gradient. Staring with 87% solvent A, consisting of 0.1% solution of NH_4OH in water, and 13% solvent B (acetonitrile). The percentage of solvent B increased to 50% in the first 10 minutes of the separation and it is kept as such for 1 minute. In the next stage the solvent B was ramped up to 95% in two minutes and kept the same for 0.5 minutes. In the final minute of the chromatogram the gradient was brought back to the initial conditions. A flow rate of 0.4 mL/min was employed during the 13.50 minutes chromatograms.

The UPLC system was coupled to an Xevo G2-S Q-TOF-MS (Waters Milford, MA, US) time of flight high resolution mass spectrometer. The

Mass spectrometer was operated with a nominal mass resolution of 35,000 and a sampling frequency of 2.3 Hz. This system was equipped with electron spray ionization source (ESI) operated in negative mode. During each cycle the mass spectrometer acquired a full-scan spectrum between 60 Da and 600 Da employing a collision energy of 6 eV.

All the samples including the blanks and quality control/assurance were analysed using the above instrumental conditions.

2.2.2.3 Quality Control/Assurance (QC)

For the purpose of QC, all the glassware used in this study were baked at 450 C overnight. The samples were divided into sets of three extracts, which were followed by a solvent injection to avoid the carryover from previous injections. Additionally, the signal of the injection standard (i.e. diazepam-D5) was monitored in order to assess the stability of the instrument during the analyses. We observed less than 20% variability in the signal of the injection standard. This suggested that all the samples showed similar levels of ion suppression for the injection standard. Therefore, we interpreted that the chromatograms were adequate for our data processing workflow without any correction for the ion suppression.

2.2.2.4 Data Analysis

All the chromatograms, including the samples and blanks, went through the following data processing steps sequentially. The acquired chromatograms were converted to an open MS format (i.e. netCDF) employing DataBridge provided via MassLynx (Waters, Milford, the US). The converted data were imported into the Matlab [21] environment (Matlab R2015b) for further processing. The imported data were mass calibrated prior to evaluation for the NAs. The details of the mass calibration are reported elsewhere [22–25]. In short, for the mass calibration, the measured mass of the calibrant injected into the source in 20 S intervals were compared to the exact mass of the same compound. The observed mass errors were used to calculate the needed mass shift over the whole chromatogram using a third order polynomial. The estimated mass shift then was applied to the data in order to produce the calibrated chromatograms. The mass calibrated data were used for the identification and signal extraction of NAs.

2.2.2.5 Identification and Signal Extraction

Each NA in a PW sample is representative of the mixture of all the structural isomers with the same molecular formula. An increase in the size of the NAs (i.e. the number of carbons) is exponentially correlated with the number of potential structural isomers of NAs [1-2]. Consequently, in the literature, NAs are typically considered as a group of isomers rather than individual compounds [2]. Similarly to the previous reports, we employed the mixture of isomers approach rather than individual compound ones.

In order to identify the NAs in our samples, a list of NAs using their general formula (i.e. $C_nH_{2n-z}O_2$) was generated. In this list the number of carbons (i.e. n) ranged between 8 to 35 while the number of rings ranged from zero to 6 (i.e. z= 0 : -2 : -20). This range was selected based on the previously reported analyzable range of NAs via LC-HRMS [2]. In addition to these conventional NAs, we added several sulfur containing NAs based on the literature reports [26], which enabled us to produce a comprehensive list of detectable NAs in PW. This resulted in a total of 181 NAs to be screened for in the samples (Appendix 3). For the identification of NAs, we generated the extracted ion chromatogram (XIC) of each NA in the list, employing a mass accuracy of ± 3 mDa. This mass window was selected based on the observed mass resolution measured using the signal of the calibrant. The generated XICs were integrated over the whole chromatogram to produce the signal specific to each NA in the list. This procedure was carried out for all the calibrated chromatograms including the blanks. The signal of each NA after the blank subtraction was used for the comparison of the performance of the three extraction methods employed in this study. During the identification, we performed a noise removal step which consisted of elimination of the NAs that produced a signal smaller than 500 counts and the NAs that were detected only in one out of three replicates. These eliminated NAs were considered non-detects for that method. This approach enabled us to accurately detect the tested NAs and compare the three extraction methods investigated in this study.

2.2.2.6 Relative Recovery Calculations

We calculated the relative recovery of each NA using the approach proposed by Samanipour et al. [18]. This approach was selected due to the large number of NAs analysed and the lack of analytical standards for individual NAs in the sample [1-2,13,16]. As an example, for an NA with formula of $C_{10}H_{18}O$ there is need for more than 37 individual analytical standards in order to define the absolute recovery of that NA. Therefore, we used the cumulative signal approach where the signal of all possible isomers of one NA is summed up to define the produced signal for that NA via an extraction method. Each NA, in this study, resulted in 9 cumulative signal values (i.e. the integrated XIC for each extract 3 methods × 3 replicates) generated via three different extraction methods. The largest method averaged cumulative signal was considered the total extractable material for that NA. Therefore, the recovery of each NA was calculated based on its signal from each extract divided by the total extractable material for that NA. Using this approach we were able to evaluate the performance of different extraction methods for each NA.

2.7 Statistical Analysis

In order to further evaluate the performance of the three extraction methods, we performed both uni-variate and multi-variate statistical analysis. For the uni-variate test, we employed the non-parametric test Kruskal-Wallis [27]. A $\rho < 0.05$ was selected as the threshold for the rejection of null-hypothesis with 95% confidence interval. With regards to multi-variate test, principal component analysis (PCA) was used in this investigation [28]. Prior to our PCA analysis our data was scaled utilizing Pareto scaling [29]. This approach has shown to be effective in keeping the data structure intact while reducing the importance of large signals. For the PCA, the singular value decomposition (SVD) was employed in order to isolate the statistically relevant components [30]. This algorithm (i.e. SVD) is effective in dealing with datasets where the number of variables is larger than the number of observations. This procedure was previously shown to be effective in separating different extraction methods from each other while isolating the variables that were causing the separation [25].

2.2.3 Results and Discussions

2.2.3.1 Detection of NAs

The ENV method with 134 positive detections out of 181 total tested NAs, performed the best, when looking at the number of positively detected NAs in the samples via different extraction methods. The HLB and Lq methods resulted in positive detection of 108 and 81 NAs, respectively (Fig. 2.2.1).

We further examined the effect of the number of rings and the number of carbons on the detection frequency of NAs produced via each extraction method.

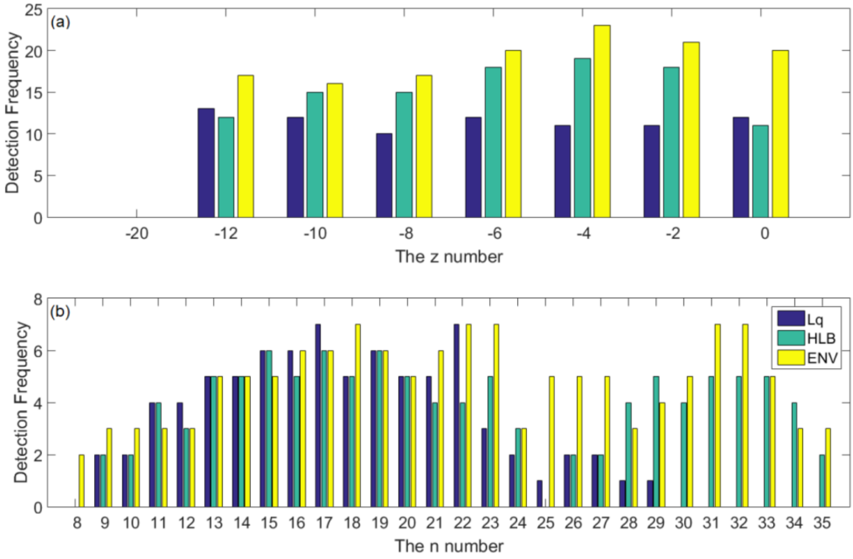

Figure 2.2.1: showing the detection frequency of NAs versus (a) the z value (i.e. the number of aliphatic rings) and (**b**) the n number (i.e. the number of carbons).

The ENV method systematically produced larger detection frequencies for all 7 z values when compared to the other two methods, Fig. 2.2.1. The largest detection frequency for both ENV and HLB was observed for NAs with a z value of -4 (i.e. 2 rings) with positive detection of 23 and 19 NAs, respectively. On the other hand, the Lq method showed to be unaffected by the number of rings in terms of the detection frequency resulting in an average of 11 NAs detected for all seven cases. The non-parametric Kruskal-Wallis test [27] results (i.e. $\rho < 0.05$) indicated that the differences observed in the detection frequencies versus the ring number were statistically significant. Further examination of these results suggested that the two SPE methods performed in a similar way whereas the Lq method appeared to be different from those two. Overall, all three methods covered a range of NAs from aliphatic chains (i.e. z=0) up to 6 rings (i.e. z=-12) while all three methods were unable to detect NAs with larger number of rings, thus z values between -14 and -20. Furthermore, none of the methods detected the sulphur containing NAs, which may suggest their absence and/or lower than instrumental limit of detection concentrations in the analysed sample.

For the effect of the number of carbons on the detection frequency of NAs, the ENV method covered all n values ranging from 8 to 35, Fig. 2.2.1. The HLB method produced zero positive detection for n values of 8 and 25 while the Lq method was limited in an n value range of 9-29. The ENV method resulted in the largest detection frequency of NAs for 20 out of 27 n values across the tested range. For cases where Lq method was the best performing approach with n values of 11, 12, 15, and 17, the mentioned NAs appeared to be aliphatic NAs. Moreover, they all were removed during the noise removal (i.e. their signal was smaller than 500 counts). For the remaining three cases with n values of 28, 29, and 34, HLB method performed better than ENV extraction method. For these cases, the missing NAs were: a one ring NA for the n value of 28, a two ring NA for the n value of 29, and finally, a five ring NA for the n of 34. Also for these cases, the noise removal step caused the elimination of these NAs from the detection list of ENV. Based on the fact that all these discrepancy cases where generated during the noise removal step, we interpreted that the sample complexity/matrix effect was the main cause of these observations. Finally, we preformed the non-parametric Kruskal-Wallis test to evaluate the trend observed in the detection frequency versus the n values. The $\rho < 0.05$ of this test suggested a statistically significant difference between the methods. Further investigation in the outcome of this statistical test showed the similarity of the SPE methods when compared to the Lq method.

Overall, the ENV method appeared to perform the best by extracting the largest number of NAs across all the z values and n values. Additionally, this method showed a consistent performance when looking at the z and n values compared to the other two methods (i.e. HLB and Lq).

2.2.3.2 Extraction Recoveries

The ENV method resulted in an average relative recovery of 49.6 % across all the tested NAs whereas HLB and Lq produced average relative recoveries of 44.7% and 42.1%, respectively. We also evaluated the recoveries of the NAs for each method based on the number of carbons and the number of rings.

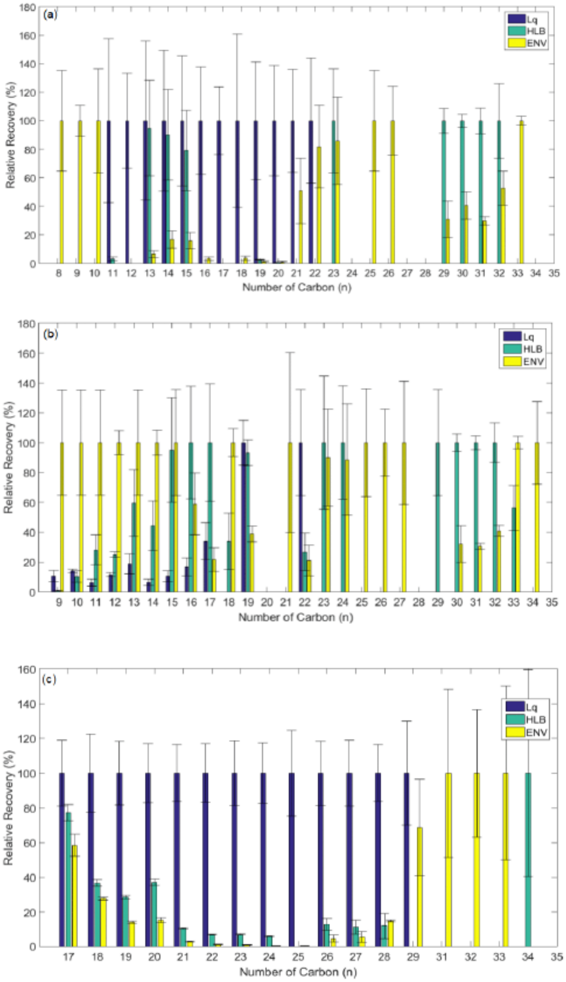

Figure 2.2.2: showing the relative recoveries of NAs versus the n value for (**a**) the z=0 (i.e. no ring), (**b**) the z=-4 (i.e. two rings), and (**c**) the z=-12 (i.e. six rings).

For the aliphatic NAs (i.e. z=0), the Lq method performed better than the other two methods resulting in 100% relative recoveries for 12 out of 27 NAs, Fig. 2.2.2. The other two methods (i.e. HLB and ENV) produced a larger level of variability in the relative extraction recoveries across the analyzed NAs, ranging from non-detect for n=12 and 17 to 100% for n larger than 29. However, the ENV method was the only method that extracted the largest number of NAs compared to the other two methods. Additionally, this method showed to be successful in capturing the smallest and the largest NAs in this group.

For small NAs with n ranging from 8 to 10 both HLB and Lq resulted in zero recoveries, which was attributed to the low affinity of these NAs for HLB resin and DCM. However, further structural elucidation is necessary to confirm this hypothesis. On the other hand, for NAs having n values larger than 22, the two SPE methods were able to isolate those NAs while the Lq failed in this task. This trend was associated with the lower solubility of larger NAs in DCM. However, in this case also further structural elucidation is necessary to confirm this hypothesis. For NAs with z values between -2 and -10 (i.e. 1 to 5 rings), the ENV method systematically produced higher relative recoveries compared to the other two methods, Fig.s 2.2.2 and 2.2.3. Among these cases, for z values of -2, -4, and -6 both ENV and Lq performed better than HLB in extracting smaller NAs. However, for NAs with n values larger than 22 the two SPE methods perform better both in terms of number of detected NAs and the relative recovery of individual NAs. Finally, for NAs with a z value of -12, thus 6 rings, the Lq performs better than the other two methods producing 100% relative extraction recoveries for 13 out of 17 NAs, Fig. 2.2.2. This method however was unable to isolate the NAs with number of carbons larger than 31. Overall, none of the methods were able to extract all the tested NAs. However, the ENV method appeared to perform better than the other two methods when looking at the relative recoveries and the number of detected of NAs.

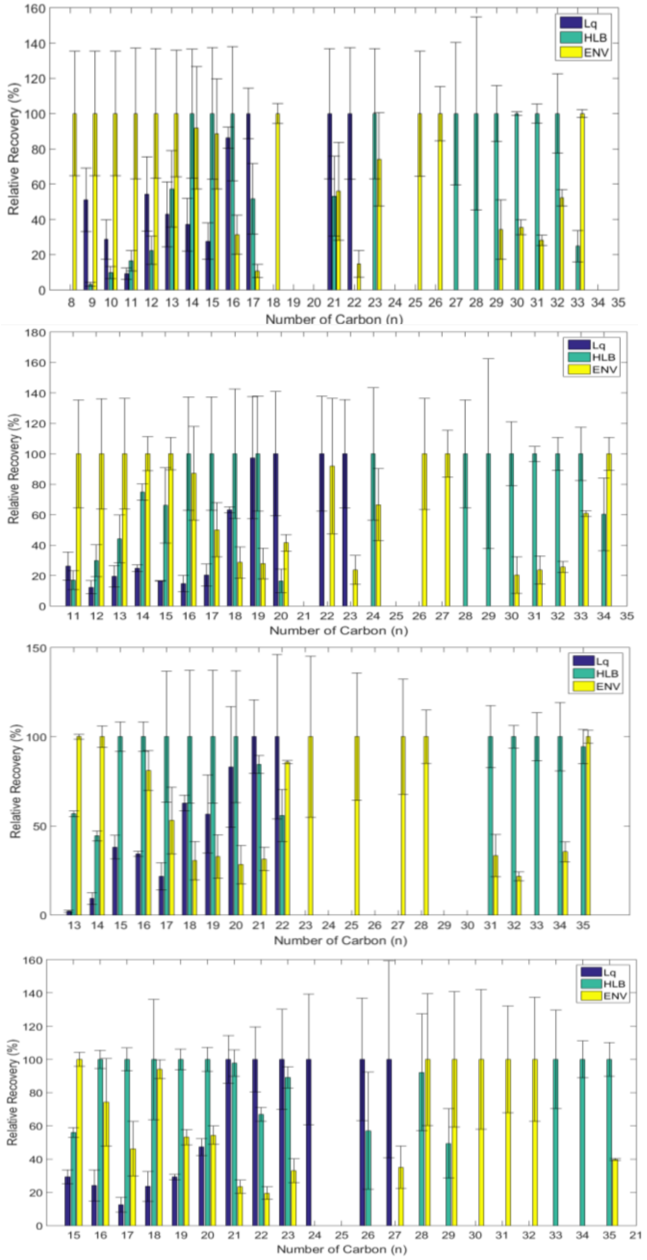

Figure 2.2.3: showing the relative recoveries of NAs with (a) the z=-2, (b) the z=-6 , (c) the z=-8 and (d) the z=-10 for all three extraction methods. The error bars show the variance observed in the data for each NA via each extraction method.

The PCA of the scaled and mean cantered relative recoveries was able to clearly distinguish the three extraction methods from each other, Fig. 2.2.4.

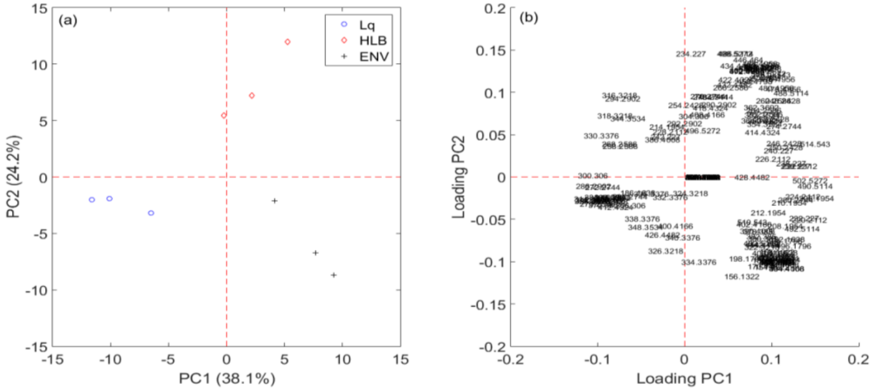

Figure 2.2.4: depicting the principal component analysis (PCA) of the scaled and mean centered

The first two PCs successfully described 62% of variability in our dataset. When looking at the loading plot, also in this case three different clusters of variables were observed. These clusters indicated the variables that were causing the separation of the methods from each other. When looking at the loadings plot, we focused on the variables that had a weight value of larger than 30%, which reduced the number of relevant variables to 79 rather than 172. From those 79, 41 were associated with the NAs where the ENV method performed better than the other two whereas 34 belonged to the method HLB. For the Lq method, there were only four statistically relevant variables (i.e. NAs with masses of 326.3218, 338.3376, 348.3534, and 426.4482), which indicated the worse performance of this method compared to the other two extraction approaches. The results of PCA suggested that the ENV method performed the best when compared to the other two methods. This was in agreement with our assessment of the recoveries based on individual NAs explained in details above.

The ENV method also produced the largest total signal of NAs compared to the other two methods, Fig. 2.2.5. We also evaluated the blank subtracted and injection standard normalized total signal of all detected NAs using each extraction method in order to evaluate the overall recovery of each method. Based on the absolute signal, the Lq and HLB methods extracted ~80% of total extractable material, assuming the ENV method extracting 100%. The outcome of the total signal was comparable to the previous reports for Lq and SPE methods [13].

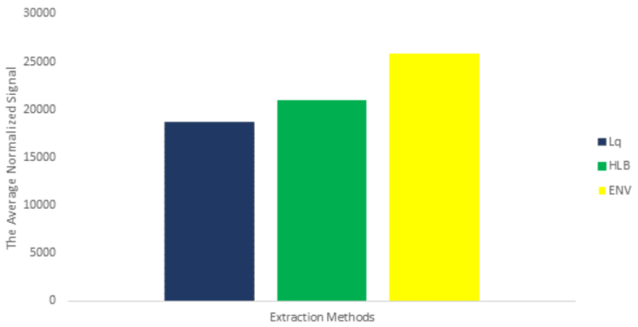

Figure 2.2.5: showing the blank subtracted and injection standard normalized total signal of all detected NAs using each extraction method.

2.2.3.3 NA Distribution in Produced Water

We further evaluated the effect of the extraction method on the overall distribution of tested NAs in the analyzed produced water. The noise removed extracted signal of the NAs for each extraction method was utilized for these evaluations.

When looking at the distribution of NAs in the analyzed produced water via SPE methods, the NAs with z values ranging from -4 to -12 appeared to be the most abundant ones. On the other hand, via Lq method the NAs with z value of -12 were the most abundant group while for other z values, this method produced relatively similar abundances, Fig. 2.2.6.

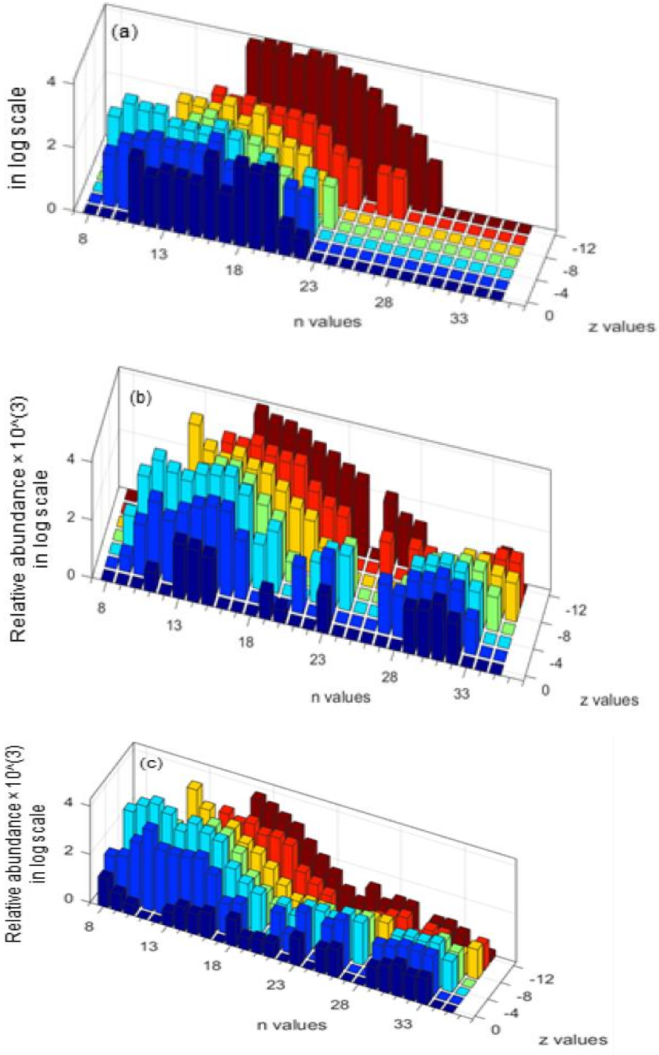

Figure 2.2.6: depicting the relative abundance of the analyzed NAs using (**a**) Lq, (**b**) HLB, and (**c**) ENV extraction methods. The relative abundances (i.e.”Z” axis) are multiplied to 1000 and are shown in log scale for ease of visual comparison among the three extraction methods.

All three extraction methods produced the smallest relative abundances for the aliphatic NAs. All the methods, for z values between -2 and -10, resulted in higher relative abundances for n values between 13 and 18, which was in agreement with previous reports regarding the distribution of NAs in produced water or similar matrices [9,31,32]. For a z value of -12, the most abundant NAs were those with n values between 16 and 20 for all three tested extraction methods.

The ENV method appeared to cover the largest NA chemical space compared to the other two methods, where the chemical space is defined as the total number of tested NAs, Fig. 2.2.6. The performance of the other SPE method, thus HLB, appeared to be more similar to the ENV rather than the Lq method. For Lq method the distribution of the NAs appeared to be affected mainly by their solubility in DCM. As a consequence, the boundaries of the explored chemical space via Lq method were dominated by the molecular size. In other words, the non-extracted NAs via the Lq were either too small or too large, therefore non soluble in DCM. For the two SPE methods, the explored chemical space appeared to be less concise when compared to the Lq method. We interpret that this observed trend was mainly caused by the interactions of individual compounds with the resin, sample complexity, and the matrix effects. We observed that the HLB method, in particular, showed less affinity for the smaller NAs (i.e. n value of 8) compared to the ENV method. To further test this, we explored our chromatograms for NAs with z value of 0 and n values of 7 and 6, which were not included in our initial list of NAs. None of the three tested extraction methods detected the NA with z=0 and n=7. However, for NA with z=0 and n=6, the ENV method was the only one producing a positive detection for that particular NA, Fig. S6. This further indicated the difficulties that the Lq and HLB methods have in extracting smaller NAs.

The ENV method was able to explore the largest chemical space of NAs compared to HLB and Lq methods. Additionally, this method was the only method that produced a positive signal for hexanoic acid, which is considered the marker for the presence of NAs in produced water according to Norwegian Oil and Gas.33 Even though this method (i.e. ENV) did not produce the highest recoveries for all the tested NAs, it resulted in 100% relative recoveries for the largest number of NAs explored in this study. Our results in overall suggested that among the tested extraction procedures the ENV method is the most effective one for analysis of NAs in produced water. However, testing the other extraction procedures is necessary and will be subject of our future study.

2.2.4 Environmental Implications

Our results suggested that the choice of sample preparation approach may have a substantial effect on the explored chemical space of NAs. In other words, using different extraction methods may produce different toxicity profiles for the same sample. This is highly relevant for a complex mixture such as produced water and NAs with a wide variety of toxicity profiles. Consequently the risk assessment of such mixtures without a comprehensive understanding of the explored chemical space becomes impossible. Our results indicated that, when dealing with such complex mixture, the conventional methods may fall short and thus the uses of more comprehensive methods are warranted. Additionally, our results indicated that when assessing the extraction recoveries, this should be done at higher detailed levels rather than the total NAs or using only a few surrogates. For example for an NA with n=24 and z=-2, this NA was detected using only one extraction method ENV, which implied that using the other two methods would not have produced an accurate toxicity profile. This is extremely important when performing the risk assessment of such complex mixtures such as NAs and PW.

Acknowledgments

We are grateful to the Research Council of Norway for the financial support of this study through project RESOLVE (grant number 243720). The authors are also thankful to equinor (former StatOil) for providing us the produced water samples.

2.2.5 References

1. Tomczyk, N.; Winans, R.; Shinn, J.; Robinson, R. On the nature and origin of acidic species in petroleum. 1. Detailed acid type distribution in a California crude oil. Energy & Fuels 2001, 15, 1498–1504.
2. Clemente, J. S.; Fedorak, P. M. A review of the occurrence, analyses, toxicity, and biodegradation of naphthenic acids. Chemosphere 2005, 60, 585–600.
3. Thomas, K. V.; Langford, K.; Petersen, K.; Smith, A. J.; Tollefsen, K. E. Effectdirected identification of naphthenic acids as important in vitro xeno-estrogens and anti-androgens in North Sea offshore produced water discharges. Environ. Sci. Technol. 2009, 43, 8066–8071.
4. Rowland, S. J.; Scarlett, A. G.; Jones, D.; West, C. E.; Frank, R. A. Diamonds in the rough: identification of individual naphthenic acids in oil sands process water. Environmen. Sci. Technol. 2011, 45, 3154–3159.
5. McCormack, P.; Jones, P.; Hetheridge, M.; Rowland, S. Analysis of oilfield produced waters and production chemicals by electrospray ionisation multi-stage mass spectrometry (ESI-MSn). Wat. Res. 2001, 35, 3567–3578.
6. Allard, A.-S.; Hynning, P.-Å.; Remberger, M.; Neilson, A. H. Environmental hazard assessment: A laboratory approach to reality? Toxicol. Environ. Chem. 2000, 78, 127–197.

7. Jones, D.; Scarlett, A.; West, C.; Frank, R.; Gieleciak, R.; Hager, D.; Pureveen, J.; Tegelaar, E.; Rowland, S. Elemental and spectroscopic characterization of fractions of an acidic extract of oil sands process water. Chemosphere 2013, 93, 1655–1664.
8. Gosselin, P.; Hrudey, S. E.; Naeth, M. A.; Plourde, A.; Therrien, R.; Van Der Kraak, G.; Xu, Z. Environmental and health impacts of Canada’s oil sands industry. Royal Society of Canada Ottawa, Ontario, Canada 2010,
9. Swigert, J. P.; Lee, C.; Wong, D. C.; White, R.; Scarlett, A. G.; West, C. E.; Rowland, S. J. Aquatic hazard assessment of a commercial sample of naphthenic acids. Chemosphere 2015, 124, 1–9.
10. Scarlett, A.; Reinardy, H.; Henry, T.; West, C.; Frank, R.; Hewitt, L.; Rowland, S. Acute toxicity of aromatic and non-aromatic fractions of naphthenic acids extractedfrom oil sands process-affected water to larval zebrafish. Chemosphere 2013, 93, 415–420.
11. Scarlett, A. G.; West, C. E.; Jones, D.; Galloway, T. S.; Rowland, S. J. Predicted toxicity of naphthenic acids present in oil sands process-affected waters to a range of environmental and human endpoints. Sci. Tot. Environ. 2012, 425, 119–127.
12. Jones, D.; Scarlett, A. G.; West, C. E.; Rowland, S. J. Toxicity of individual naphthenic acids to Vibrio fischeri. Environ. Sci. Technol. 2011, 45, 9776–9782.
13. Kovalchik, K. A.; MacLennan, M. S.; Peru, K. M.; Headley, J. V.; Chen, D. D. Standard method design considerations for semi-quantification of total naphthenic acids in oil sands process affected water by mass spectrometry: A review. Front. Chem. Sci. Eng. 2017, 11, 497–507.
14. Barros, E. V.; Dias, H. P.; Pinto, F. E.; Gomes, A. O.; Moura, R. R.; Neto, A. C.; Freitas, J. C.; Aquije, G. M.; Vaz, B. G.; Romaao, W. Characterization of Naphthenic Acids in Thermally Degraded Petroleum by ESI (-)-FT-ICR MS and 1H NMR after Solid-Phase Extraction and Liquid/Liquid Extraction. Energy & Fuels 2018, 32, 2878– 2888.

15. Huang, R.; McPhedran, K. N.; Sun, N.; Chelme-Ayala, P.; El-Din, M. G. Investigation of the impact of organic solvent type and solution pH on the extraction efficiency of naphthenic acids from oil sands process-affected water. Chemosphere 2016, 146, 472–477.

16. Huang, R.; Chen, Y.; Meshref, M. N.; Chelme-Ayala, P.; Dong, S.; Ibrahim, M. D.; Wang, C.; Klamerth, N.; Hughes, S. A.; Headley, J. V. Monitoring of classical, oxidized, and heteroatomic naphthenic acids species in oil sands process water and groundwater from the active oil sands operation area. Sci. Tot. Environ. 2018, 645, 277–285.

17. Mohammed, M. A.; Sorbie, K. S. Naphthenic acid extraction and characterization from naphthenate field deposits and crude oils using ESMS and APCI-MS. Colloids and Surfaces A: Physicochemical and Engineering Aspects 2009, 349, 1–18.

18. Samanipour, S.; Baz-Lomba, J. A.; Reid, M. J.; Ciceri, E.; Rowland, S.; Nilsson, P.; Thomas, K. V. Assessing sample extraction efficiencies for the analysis of complex unresolved mixtures of organic pollutants: A comprehensive non-target approach. Anal. Chimica Acta 2018, 1025, 92–98.

19. Martin, J. W.; Han, X.; Peru, K. M.; Headley, J. V. Comparison of high-and lowresolution electrospray ionization mass spectrometry for the analysis of naphthenic acid mixtures in oil sands process water. Rapid Commun. Mass Spectrom. 2008, 22, 1919–1924.

20. Statoil, N. Heidrun oil platform. https://www.statoil.com/en/what-we-do/norwegian-continental-shelf-platforms/heidrun.html, 2017.

21. MATLAB version 9.1 Natick, Massachusetts: The MathWorks Inc., 2016.

22. Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Results. Environ. Sci Technol. 2018, 52, 4694–4701.

23. Samanipour, S.; Baz-Lomba, J. A.; Alygizakis, N. A.; Reid, M. J.; Thomaidis, N. S.; Thomas, K. V. Two stage algorithm vs commonly used approaches for

the suspect screening of complex environmental samples analyzed via liquid chromatography high resolution time of flight mass spectroscopy: A test study. J. Chromatogr. A 2017, 1501 (2017), 68–78.

24. Samanipour, S.; Langford, K.; Reid, M. J.; Thomas, K. V. A two stage algorithm for target and suspect analysis of produced water via gas chromatography coupled withhigh resolution time of flight mass spectrometry. J. Chromatogr. A 2016, 1463, 153–161.
25. Samanipour, S.; Reid, M. J.; Thomas, K. V. Statistical variable selection: An alternative prioritization strategy during the non-target analysis of LC-HR-MS data. Anal. Chem. 2017, 89 (10), 5585–5591.
26. Qian, K.; Robbins, W. K.; Hughey, C. A.; Cooper, H. J.; Rodgers, R. P.; Marshall, A. G. Resolution and identification of elemental compositions for more than 3000 crude acids in heavy petroleum by negative-ion microelectrospray high-field Fourier transform ion cyclotron resonance mass spectrometry. Energy & Fuels 2001, 15, 1505–1511.
27. Breslow, N. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 1970, 57, 579–594.
28. Brereton, R. G. Applied chemometrics for scientists; John Wiley & Sons, 2007.
29. van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC genomics 2006, 7, 142.
30. Golub, G. H.; Reinsch, C. Singular value decomposition and least squares solutions. Numerische mathematik 1970, 14, 403–420.
31. Holowenko, F. M.; MacKinnon, M. D.; Fedorak, P. M. Characterization of naphthenic acids in oil sands wastewaters by gas chromatography-mass spectrometry. Water Res. 2002, 36, 2843–2855.
32. Rogers, V. V.; Liber, K.; MacKinnon, M. D. Isolation and characterization of naphthenic acids from Athabasca oil sands tailings pond water. Chemosphere 2002, 48, 519–527.

33. Norog, Norwegian Oil and Gas recommended guidelines for sampling and analysis of produced water, translated version. 2003.

Chapter 3

Chemometric Strategies in Industry Projects

3.1 Project VI

Identification of Base Stock in Engine Oils by Near Infrared and Fluorescence Spectroscopies coupled with Chemometrics

Summary

Engine oils (also called motor oils or engine lubricants) are produced by blending almost 80% (w/w) base oils (a mixture of one or more base stock types) and 20% (w/w) of different additives. The American Petroleum Institute (API) has categorized base stocks into five categories; to date, it is possible to identify the base stock type by looking at the combination of physical properties (viscosity index, density, colour, flash point, pour point, aniline point, thermal stability) but the measurement of these parameters is expensive and time consuming. The aim of the present study was to investigate, for the first time, the capabilities of near infrared (NIR) and excitation-emission (EEM) fluorescence spectroscopies coupled with chemometrics as low-cost, green and non-destructive methods in order to identify the type of base stock into engine oil. In order to reach this goal, 53 pure base stocks belonging to different API groups and 43 engine oils were analysed without any pre-treatments. PCA performed on the NIR and EEM unfolded spectra showed that samples form clusters according to their API groups and to their chemical composition. Considering the 3-ways nature of the EEM data, PARAFAC was also applied on fluorescence data and outcomes were in agreement with PCA results. PLS-DA, as a multivariate classification tool, was applied in order to distinguish among different API base stock groups and satisfactory results were achieved: the prediction abilities in the external test set were 87% and 85% using NIR and

fluorescence spectroscopy, respectively. Moreover, in the present study, the performance level of gasoline engine oils, as a recognition aspect of lubricants quality, was also investigated. Both spectroscopic techniques appeared to be rapid and non-destructive analytical methods for the characterization of base stock and for the determination of the performance level, therefore, they represent a promising tool for engine oil analysis.

3.1.1 Introduction

Engine oil, also called motor oil has the largest consumer market among lubricants and is facing a new challenge every day to meet the new demands of the automotive industry such as environmental constraints and the need to save energy. Manufacturers of lubricants, package producer and standard institutes are dealing with the costly and time-consuming processes of producing, standardizing and modifying based on new vehicle developments. On the other hand, the huge turnovers of this field have fuelled scams, counterfeits, frauds, tax evasion up to low quality oils that cause irreversible damage to engines.[1]

Engine oil is produced by blending almost 80% (w/w) base oils (a mixture of one or more base stock types) and 20% (w/w) of different additives to meet the performance level requirements. Since base oil is the major part of the lubricant formula, and additives are only added to improve the base oil performances, the identification of the base stock type can, provide a reasonable estimation of engine oil quality.

The American Petroleum Institute (API) is the largest U.S. trade association for the oil and natural gas industry. API has categorized base stocks into five categories. The first three groups are mineral stocks, refined from crude oil with different severity processes on lub-cut (middle cut of vacuum distillation tower in crude oil refinery). Group IV base stocks are full synthetic (polyalphaolefin). Group V is for all other base stocks not included in Groups I through IV. (See Table 3.1) [2]

Table 3.1: API Base Stock Categories

	Base Stock Category	**Sulfur (%)**	**Saturation (%)**	**VI**
Mineral	**Group I**	>0.03 and/or	<90	80 to 120
	Group II	<0.03 and	>90	80 to 120
	Group III	>0.03 and	>90	> 120
Synthetic	**Group IV**	Poly Alpha Olefin (PAO) Synthetic Base Stocks		
	Group V	All other base stocks not included in Group I, II, III, IV		

The properties of the three mineral base stocks are not similar due to different refining processes. Moreover, since their source is natural, their molecules have different size and structures containing various elements such as Oxygen, Phosphorous, Nitrogen and Sulphur along with Carbon and Hydrogen backbone. Producing processes only remove some undesired structures or break down some detrimental bands. On the contrary, synthetic base stocks molecules have same shape and size tailored for specific demands.

In laboratory, it is possible to distinguish base stock types by looking at the combination of physical properties such as viscosity index, density, colour, flash point, pour point, aniline point, thermal stability. Nevertheless, identification of a mixture of base stocks, especially in lubricants, represents a major analytical challenge, due to the variable composition of base stock and additives.

In the present study, the performance level of engine oils, as a recognition aspect of lubricants quality, was also investigated. API's Certification Mark and Service Symbol identify quality motor oils for gasoline and diesel powered vehicles. Oils displaying these marks meet performance requirements set by U.S. and international vehicle and engine manufacturers and the lubricant industry. [3]

The API "Donut" in Figure 3.1 identifies oils that meet current API engine oil standards. It includes the Performance level classification (see part 1 of Figure1). Regarding Performance level, the letter "S" refers to oil suitable for gasoline engines, and the letter "C" refers to oil suitable for diesel engines. “S” refers to Service/Spark Ignition (petrol) and “C” to Commercial/Compression Ignition (diesel). Letter ‘S’ is followed by another letter from “A” (first class for production cars up to 1930) to “N” (highest performance level); Letter ‘C’ is followed by another letter from “A” (first performance level) to “I” (latest level). "4" in CI-4 refers to the combustion cycle of the engine. [3]

Therefore, the latest API service category is API SN Plus for gasoline automobile engines. As an instance, the SN standard refers to a group of laboratory and engine tests, including the latest series for control of high-temperature deposits. Current API service categories include SN, SM, SL and SJ for gasoline engines. All earlier service categories are obsolete, although they are still produced and used in some parts of the world. [4]

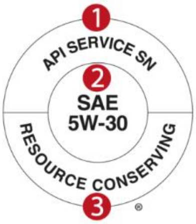

Figure 3.1: API "Donut”: **1)** This part displays the motor oil's API performance standard; **2)** The centre of the "Donut" shows the motor oil's SAE viscosity grade; **3)** The bottom tells whether the motor oil has resource-conserving properties when compared with a reference oil in an engine test.

In order to achieve the API mark and certificates, specialized laboratories around the world perform some expensive tests to cover all standard requirements for each performance level. [1]

A rapid solution in order to identify the type of base stock in engine oils could help the formulators when developing a new or tailored lubricant, targeting a given performance level. Since spectroscopic techniques are low-cost, green, non-destructive and fast, in order to reach this goal, in the present study the capabilities of Fourier Transform Near Infrared (FT-NIR) spectroscopy and EEM fluorescence spectroscopy coupled with chemometrics have been investigated in the analysis of motor oils, for the first time.

The spectrophotometers illustrate the effect of electromagnetic radiation on matter, which appears as absorption or emission intensity of electron transfer between the bending, stretching, bonding, or etc. atomic or molecular quantum layers. Therefore, since each intensity is sensitive to position and type of elements in a molecule, it is powerful way to find particular bond, group of agent or interaction in the matter.

Base stocks and engine oils are mixtures of different chemical compounds. Thus their NIR spectra are very complex and Chemometrics, as a statistical tool, is necessary to analyse this large amount of data containing a lot of information [5]. The same is for fluorescence data. Chemometrics in analysing complex chemical mixtures is a cutting edge method to look at a large amount of data containing all of its information [6].

To the best of our knowledge, few analytical methods have been proposed coupled with chemometrical techniques in order to analyse engine oils in general and to characterize base stocks according to their API group, in particular. As an instance, Poppi et al. [7] applied Fourier transform infrared (FTIR) spectroscopy in combination with multivariate statistics, based on PCA, to develop a quality control strategy for classification of lubricant type (mineral, synthetic and semi-synthetic) and usage conditions.

Amat et al. [8] presents a method to evaluate the lubricant oil oxidation using NIR coupled with chemometrics. In another study of Hirri et al. [9], FTIR spectroscopy coupled to chemometric techniques, like PLS2-DA and PCA, was reported as an adequate method for the quality control of lubricating oils SAE 30 of gear and machines in industries.

3.1.2 Material and Methods

3.1.2.1 Sample and Reagents

The sampling was possible thanks to the collaboration with three different petrochemical companies in fact, for the present study, guaranteed samples with known composition were required. Fifty-three base stock and forty-three engine oil samples were provided by the petrochemical companies in different times and analysed in two different working sessions. The complete lists of samples are reported in Appendix 4 a and 5. The base stocks belonged to the API categories from I to IV, there were not samples for categories V and VI that are the least used categories for the formulation of engine oils.

As far as engine oils are concerned, it can be seen from Appendix 5 that some oils contain only one type of base stock, others are mixtures of several base stocks.

All oils have been tested two or three times.

3.1.2.2 Instrumental and Methods

In this study, excitation-emission (EEM) fluorescence and near infrared (NIR) spectroscopies were applied as inexpensive and rapid analytical techniques in order to analyse both base stock and engine oil samples.

NIR spectra were acquired with a FT-NIR spectrophotometer (Buchi NIRFlex N-500), in the 4000-10000 cm^{-1} range at 4 cm^{-1} resolution, in transmittance mode. All the experiments were performed at controlled temperature (35℃).

EEM spectra were measured using a luminescence spectrometry (LS-55, Perkin-Elmer Co., USA). EEM spectra are a collection of a series of emission spectra over a range of excitation wavelengths, and they can be used to identify fluorescent compounds present in complex mixtures. In this study, EEM spectra were collected with subsequent scanning emission spectra from 300 to 900 nm at 0.5 nm increments by varying the excitation wavelength from 200 to 500 nm at 10 nm increments. The excitation and emission monochromator slits were set to 4.5 and 11.0 nm, respectively, and the scanning speed was set at 200 nm/min for all the measurements.

The EEM spectra were measured after optimisation of the operating parameters. In order to optimise the factors of the fluorescence spectrometer, changing one variable at a time is a most popular way, but design of experiments (DoE) [10] as a multivariate approach, represents a powerful way to reduce time and cost without losing significant information. In the present study, a D-optimal design [11] as a method of multivariate DoE, was applied. According to our recent study [12] the sample loadings on the first principal component for each experiment were the response considered in the D-optimal design.

Principal component analysis (PCA) [13] was performed as a multivariate display method in order to visualize the NIR and unfolded EEM data structure.

In order to remove the effect due to the different period of analysis, NIR spectra were block-scaled; in particular NIR spectra were pre-processed by Standard Normal Variate (SNV) to correct for light scatter and autoscaling for each block of analysis. EEM spectra were block-autoscaled before data analysis.

According to the specific nature of EEM data, organized in a 3D data array (sample × excitation × emission), for performing PCA a step of unfolding of the matrix is request while with the PARAFAC algorithm [14] is possible to model directly n-way data [15]. In the case of three-way data, PARAFAC decomposes a data tensor **X** with dimension I × J × K into three loading matrices **A**, **B** and **C**, being their columns $\mathbf{a_f}$, $\mathbf{b_f}$ and $\mathbf{c_f}$ are their column respectively. The trilinear PARAFAC model is expressed as follows:

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

where x_{ijk} is the array of **X** tensor in the position i, j, k. e_{ijk} is the residual array in the position i, j, k of residual tensor (a tensor with same dimensions of **X** tensor). *F* is the number of PARAFAC components (factors). In the case of fluorescence data, $\mathbf{a_f}$, $\mathbf{b_f}$ and $\mathbf{c_f}$ are *f*-th fluorophore (component) excitation, emission and sample profile. Sample profile refers to quantity of *f*-th fluorophore in each sample. [16]

Lastly, Partial least squares discriminant analysis (PLS-DA) [17], as a multivariate classification chemometric method, was applied in order to discriminate mineral base stocks according to the base stock API group. The

PLS-DA classification rules were validated both in cross validation and using on external test set.

PLS-DA is based on the partial least squares (PLS) [18] regression algorithm. PLS-DA is a bilinear regression method that finds the relationship between predictor variables (**X**) and dependent categorical variable (**Y**). This classification method has been used separately in NIR and fluorescence data.

The FL WinLab software (PerkinElmer) was used to register the fluorescent signals and the NIRWare 1.5 software (Buchi) was used to register the NIR spectra.

The data were imported to MATLAB [19] and PARAFAC, PCA and PLS-DA models were performed using PLS Toolbox [20].

3.1.3 Results and Discussions

3.1.3.1 NIR Spectroscopy

3.1.3.1.1 Repeatability Study

The multivariate repeatability for NIR analysis was assessed. It was made by replicating three times the analysis of three base stock samples of different API category. Figure 3.2 displays the score plot obtained by PCA of these base stock NIR data: there was not

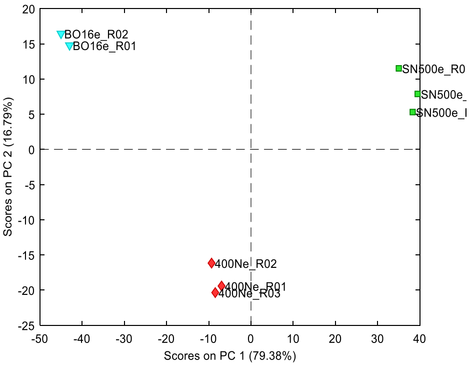

Figure 3.2: Score plot of Three Base Stock NIR Replicate

evident effect on repeatability, so the rest of data processing was performed using the average spectra.

3.1.3.1.2 Data Exploration (PCA)

Despite the calibration of the instrument, it is possible to notice a clear difference between the baselines of the NIR spectra acquired in the two different working sessions (WS) (Fig. 3.3). Figure 3.4 shows the score plot obtained on these NIR spectra after SNV and mean centering. As expected, PC1, the direction explaining the maximum variation of the data (96% of the total variace), perfectly discriminates the samples according to the two WS of analysis.

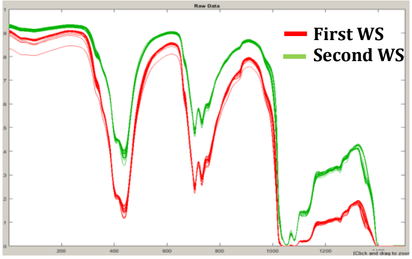

Figure 3.3: Raw NIR Spectra of Base Stocks. Each colour is related to one WS.

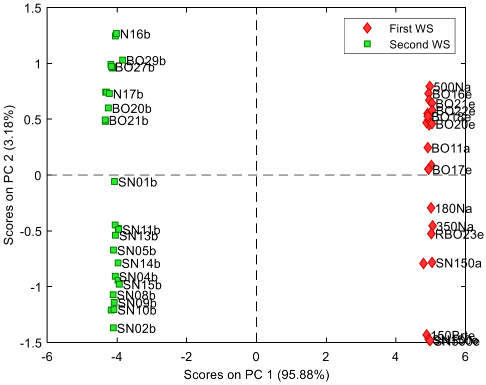

Figure 3.4: Score Plot of 53 Base Stock NIR Spectra (Pre-processing: SNV + Mean Centering)

Scaling block-wise of the NIR spectral matrix was performed in order to eliminate the effect of the 'block', in this case the difference due to the different WS of analysis, and PCA was performed again on this pre-treated matrix.

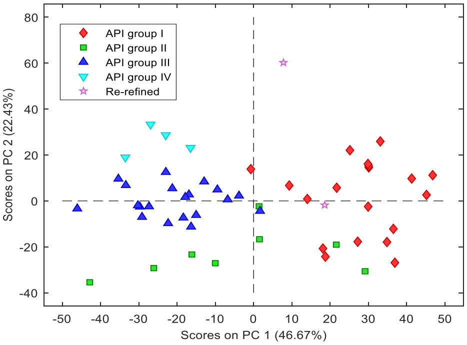

Figure 3.5: Score Plot of 53 Base Stock NIR spectra (Pre-processing: Block SNV + Auto Scaling)

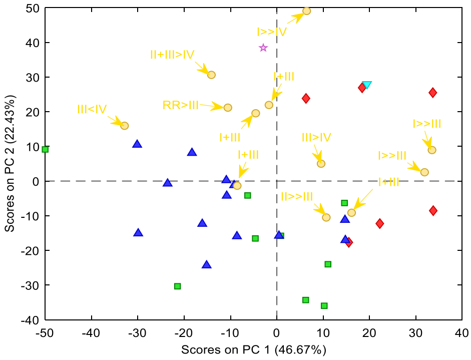

Figure 3.6: Projection Av. Engine Oils on Av. Base Stocks PCs (Block SNV + AutoScaling)

In the score plot of Figure 3.5, samples form clusters according to their API group. It is difficult to recognize if each PC describes specific chemical or physical properties of the samples, but, it appears that along PC1, base stocks have been apart because of heteroatoms existence and the saturation degree seems the reason of sample distribution along PC2.

The score plot shows high agreement with the chemical structure of the oils and it appears that heteroatoms (along PC1) have stronger effects than the saturation degree (along PC2) on varying the intensity of the NIR spectra.

Subsequently, the NIR spectra of the 43 engine oil samples were projected into the plane of the first 2 PCs calculated using the NIR spectra of the base stocks, in order to compare the position of each engine oil with respect to its base oil composition and to evaluate the pattern's similarity. It was interesting to notice that the position of the engine oil samples in this score plot was in agreement with their chemical composition (see Fig. 3.6).

Among the 43 engine oil samples provided by the petrochemical companies, only 33 gasoline engine oil samples had declared performance levels, therefore these samples were used in this second part of the study where the link between base oil composition and performance level of engine oil has been investigated. PCA was performed on these 33 engine oil samples and the PCA score plot (Fig. 3.7) was coloured in two different ways: in figure 7a, the samples are coloured based on base stock type used in engine oil and in Figure 7b according to engine oil performance level. The two score plots show approximately same trend (grey arrow) from low to high API base stock group (a) or performance level (b).

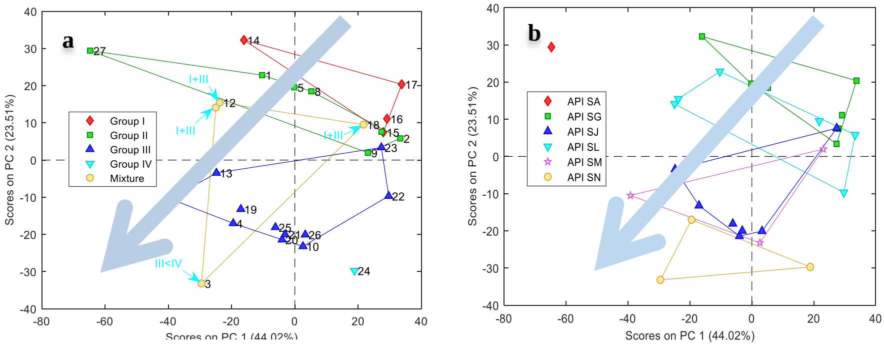

Figure 3.7: Engine Oil Coloured According to **a:** Base Stock Type and **b:** Performance Level. (Pre-processing: Block SNV + Auto Scaling)

3.1.3.1.3 Classification Analysis (PLS-DA)

Partial Least Squares Discriminant Analysis (PLS-DA) was applied as a classification technique in order to discriminate the base oil samples according to the first three API categories. Categories IV and re-refined base oil were not considered in this classification approach because the number of samples available for these 2 categories was too limited for the development of a class model.

In Table 3.1, the PLS-DA results obtained in cross-validation (5 cancellation groups) are reported and the corresponding plots are represented in Figure 3.6. In cross validation, base oils of API groups I and III are correctly predicted in class 1 and 3, respectively; on the contrary, the prediction of API group II samples was harder probably due to the lower number of samples (Fig. 3.8a-c). Red dashed lines is the best threshold estimated using Bayes Theorem [21] (number of false positives and false negatives is minimized) for each class. The sensitivity indicates the total number of correctly classified samples in the studied class and the specificity the samples of other classes correctly rejected by the class model. A correct prediction ability of 87% was achieved on an external test set composed of 20% of samples randomly selected.

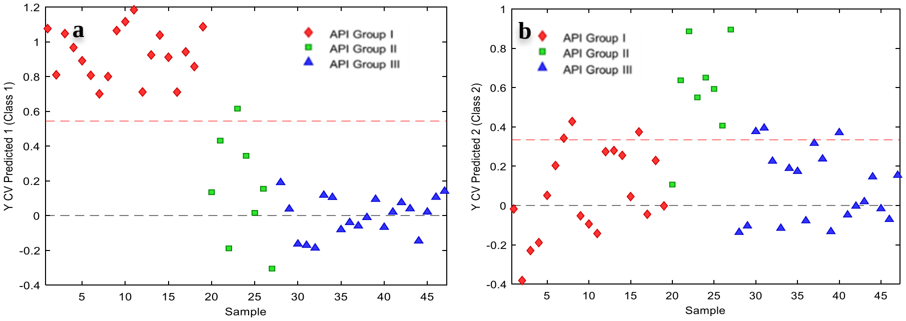

Table 3.1: PLS-DA Result of Mineral Base Oil NIR

	Class 1	Class 2	Class 3
Cal.* Sensitivity	1.00	0.87	1.00
CV Sensitivity**	1.00	0.87	1.00
Cal. Specificity	0.96	0.92	0.96
CV Specificity	0.96	0.92	0.96
Cal. Class Er.***	0.0178	0.1009	0.0185
Cal. Class Er.	0.0178	0.1138	0.0185

*Cal.: Calibration
** CV: Cross Validation
***Er.: Error

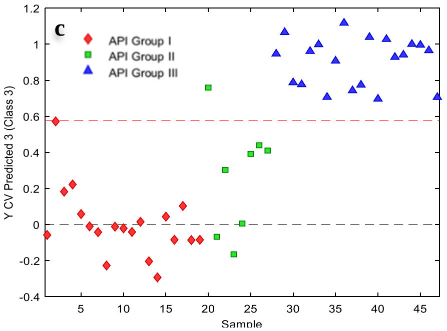

Figure 3.8: Prediction Ability in **a:** Class 1, **b:** Class 2 and **c:** Class 3 by PLS-DA

3.1.3.2 Spectrofluorimetry

3.1.3.2.1 Repeatability Study

Also for EEM spectra, the multivariate repeatability was assessed by replicating three samples for three times. Figure 3.9 displays the PCA score plot of the

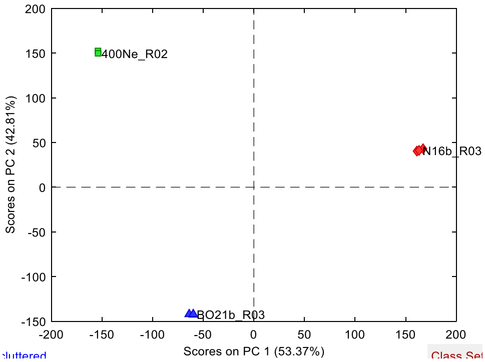

Figure 3.9: Score plot of Three Base Stock Fluorescence Replicates

EEM spectra of base stock: like for the NIR outcome, there was not evident effect on repeatability.

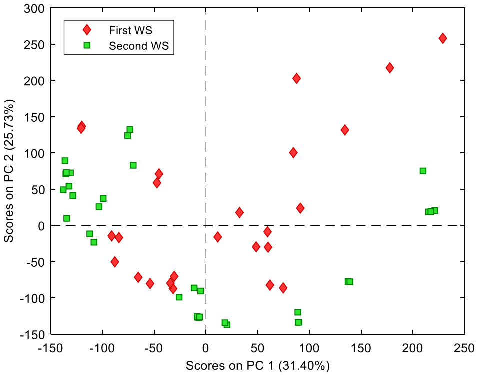

Figure 3.10: Base Oil Fluorescence Coloured According to the WS

3.1.3.2.2 Data Exploration

As for NIR data, time interval in sampling affects the fluorescence spectra even if in minor way. PCA analysis on unfolded data confirmed that fluorescence spectra were slightly affected by difference in WS of analysis (see score plot in Figure 3.10). Figures 3.11a-c, sow the score plots performed on each type of base stock separately, which demonstrate this light effect.

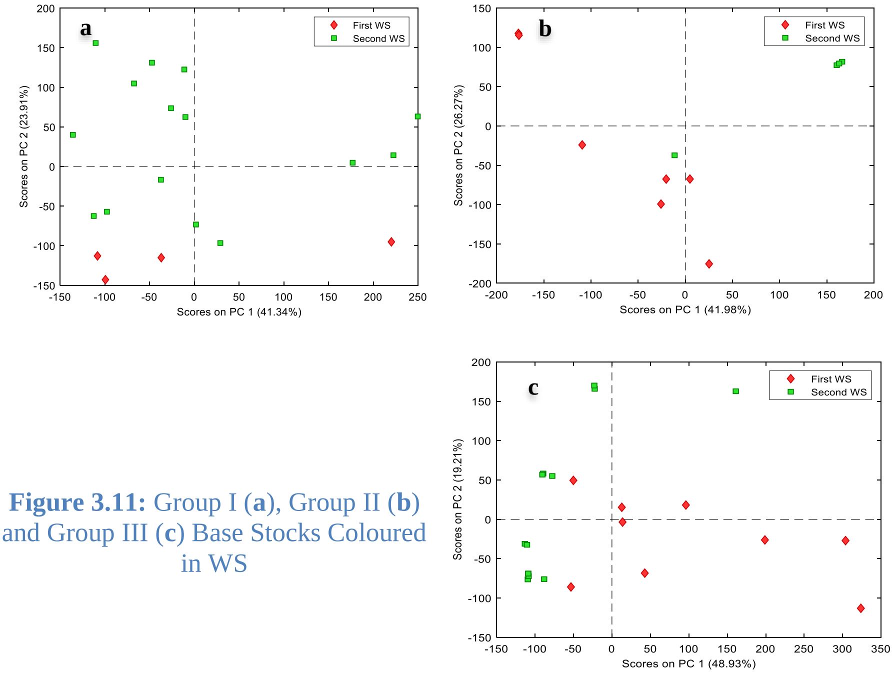

Figure 3.11: Group I (**a**), Group II (**b**) and Group III (**c**) Base Stocks Coloured in WS

In order to avoid any possible effect on the final result, the EEM data were block-autoscaled.

3.1.3.2.2.1 Two-Way Visualization (PCA)

Principal Component Analysis was used as a display method in order to visualize the unfolded EEM data structure.

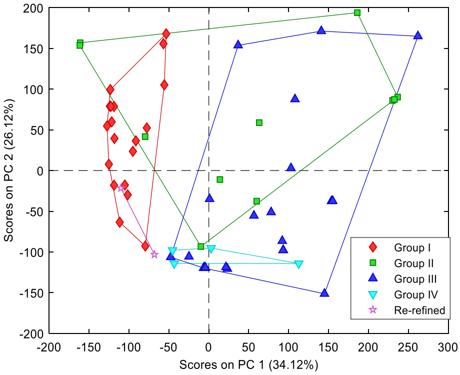

Figure 3.12: PCA of Unfolded Base Stock Fluorescence (Pre-processing: Block Auto Scaling)

PCA was initially performed on the base stocks data (Fig. 3.12) and then on the engine oils data matrix (Fig. 3.13). In figure 3.12, API groups I and III samples are completely separated, while, group II, IV and re-refined base stocks overlap with other groups. In addition, the difference between group I (and re-refining) and the other groups were identified along PC1.

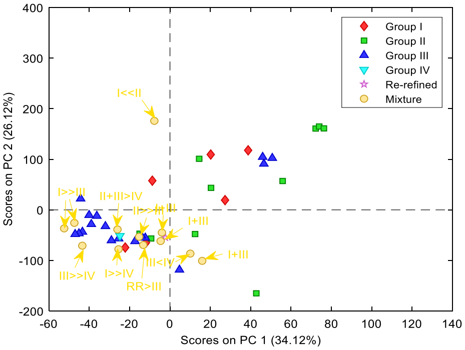

Figure 3.13: Projection Av. Engine Oils on Av. Base Stocks PCs (Block AutoScaling)

In figure 3.13 it is difficult to find a similar clustering or trend in the engine oil projection on base stock PCs, Although, considering their base oil composition, engine oils with similar base stock are closer. As for NIR data, among the 43 engine oil samples provided by the petrochemical companies, only the 33 samples with declared performance levels were used to investigate the link between base oil composition and performance level of engine oil.

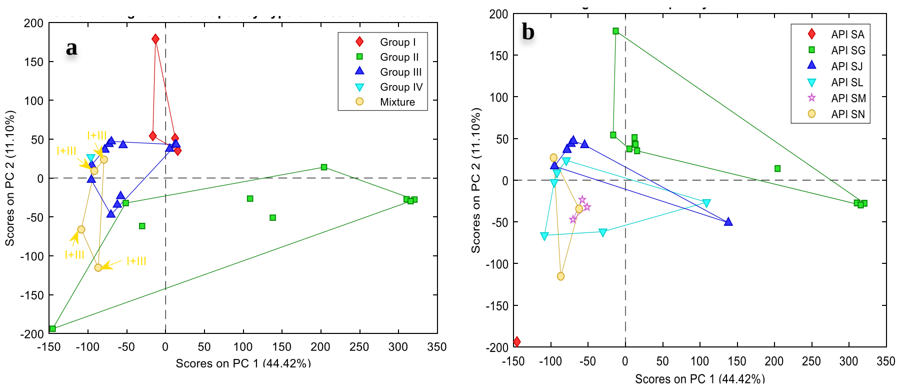

Figure 3.14: Engine Oil Coloured Based on **a:** Type of Used Base Oil and **b:** Performance Level. (Pre-processing: Block SNV + Auto Scaling)

PCA was performed on these 33 EEM spectra and the PCA score plot was coloured according to the base oil type (Fig. 3.14a) and to the performance level (Fig 3.14b). Differently from NIR, a specific and common trend was not observed.

3.1.3.2.2.2 Three-way Analysis (PARAFAC)

The EEMs recorded for the 47 mineral base stock and 43 engine oil samples with replicates analysed were arranged into data tensors (data cubes) where the excitation wavelengths between 200 nm and 500 nm and the emission wavelengths between 300 nm and 900 nm were considered. Therefore, the dimension of these tensors were 52 × 1201 × 31 (47 samples plus 5 replicates × emission × excitation) and 48 (43 samples plus 5 replicates) × 1201 × 31 respectively for base oil and engine oil samples.

The PARAFAC decomposition of these tensors required linearity in three factors (CORE CONSISTENCY [16] of 100% and 99% respectively).

To obtain high core consistency for mineral base stocks (Fig. 3.15e) and engine oil (Fig. 3.16e), PARAFAC analysis was utilized with same pre-treatment (scaled in emission mode) on both cube of data.

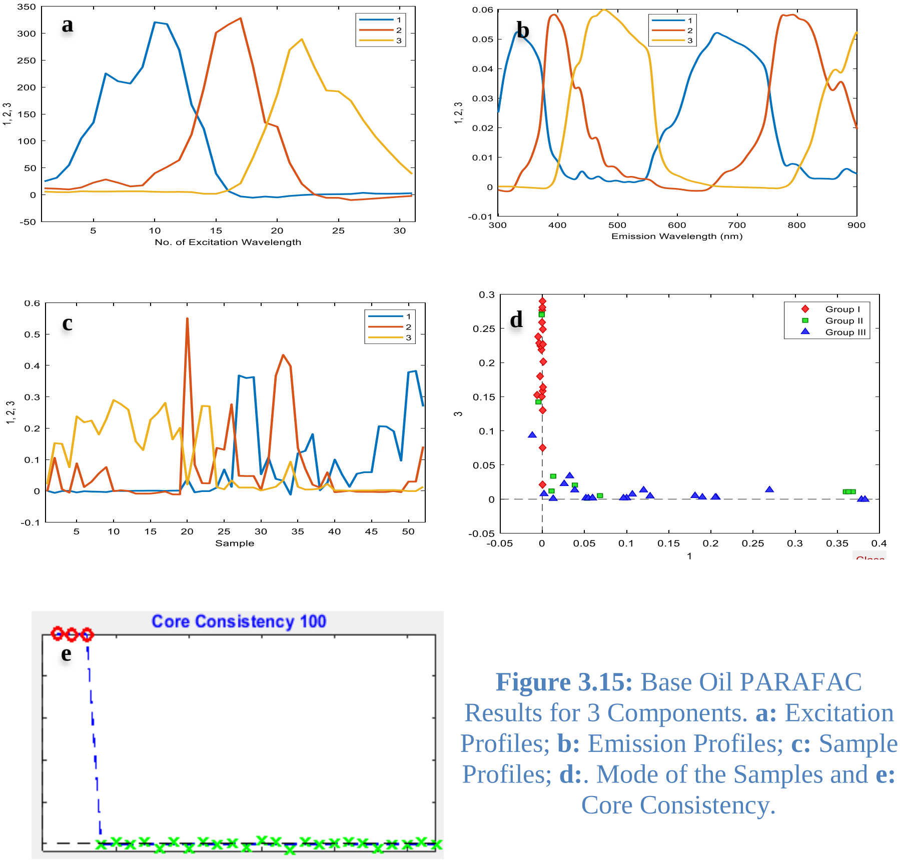

Figure 3.15: Base Oil PARAFAC Results for 3 Components. **a:** Excitation Profiles; **b:** Emission Profiles; **c:** Sample Profiles; **d:**. Mode of the Samples and **e:** Core Consistency.

In sample profile of mineral base stocks, samples from 1 to 19 are group I; from 20 to 30 are group II and from 31 to 52 are group III. According to the sample profile, group I and group III can be differentiated using two components (1 and 3). But it is difficult to use same statement about groups I and II, and, groups II and III. Based on this profile, component 2 almost represents the group II base oil. It is also illustrated in Fig. 3.15d, how all the group I are distributed along the compound 3 axis and the group III oils along the compound 1 axis . And, as observed in the PCA, group II oils have the highest overlap with group III oils.

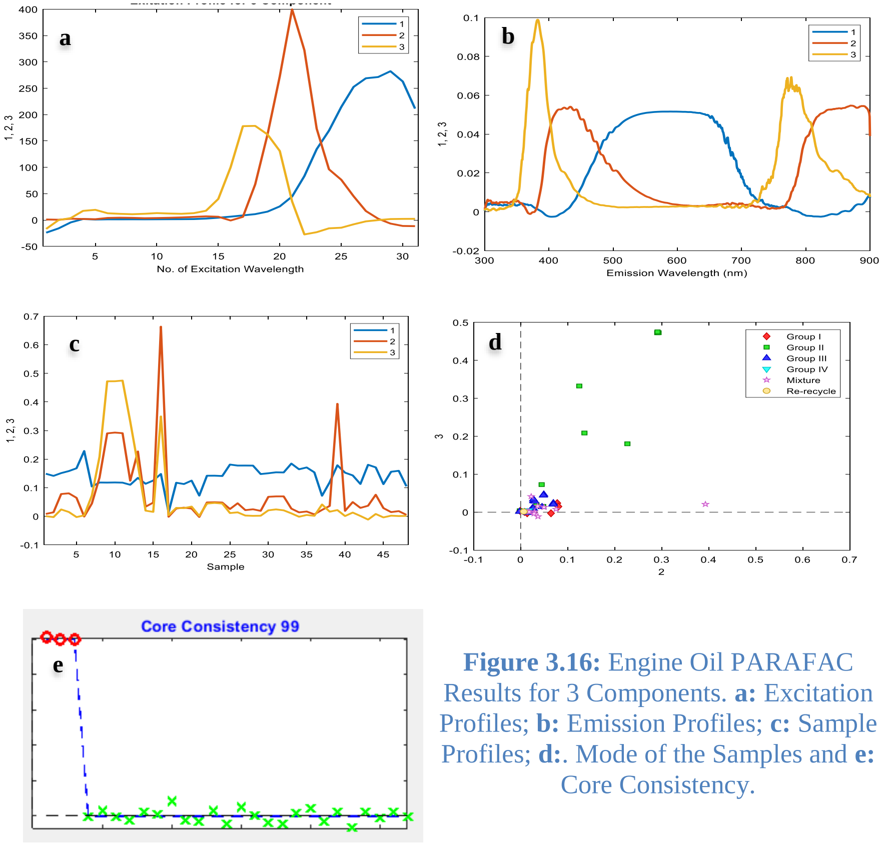

Figure 3.16: Engine Oil PARAFAC Results for 3 Components. **a:** Excitation Profiles; **b:** Emission Profiles; **c:** Sample Profiles; **d:**. Mode of the Samples and **e:** Core Consistency.

By comparing emission profiles of engine oils (Fig. 3.16b) and mineral base oils (Fig. 3.15b) it appears that component 2 in mineral oil (red line) has as same as picks of component 3 in engine oil (yellow line), and emission picks of component 3 mineral oil (yellow line) is as almost same as component 2 of engine oil (red line). But it is a little difficult to find correlation between two blue lines in two different emission profiles. On the other hand, it appears that the most effective excitation part for this recognition between base oil (Fig. 3.15a) and engine oil (Fig. 3.16a) are from 15 to 25 scan number of excitation wavelength which is related to 340 to 440 nm.

The plot of the loadings of the mode of the samples (first mode of base oil EEM, Fig. 15d is similar to the PCA score plot (Fig. 3.12), and shows a clear discrimination between API group I and Group III samples.

3.1.3.2.3 Classification Analysis (PLS-DA)

In the case of classification, according to the base oil API categories, PLS-DA on mineral base oil data predicted correctly 85% of the external test set composed of 20% of randomly selected samples. The result of discrimination analysis was roughly similar to the NIR outcome. (Table 3.2 and Fig. 3.17a-c)

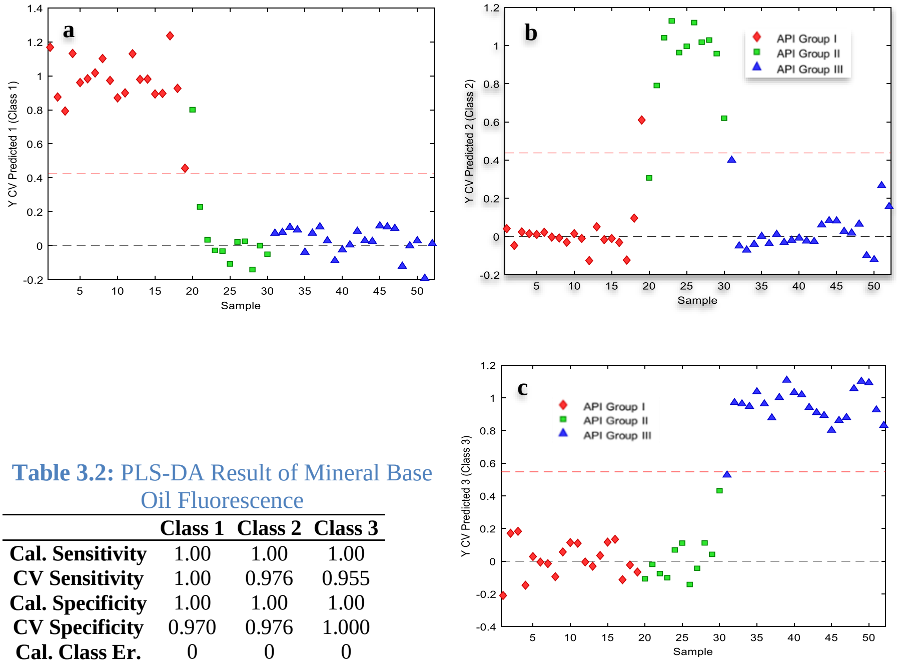

Table 3.2: PLS-DA Result of Mineral Base Oil Fluorescence

	Class 1	Class 2	Class 3
Cal. Sensitivity	1.00	1.00	1.00
CV Sensitivity	1.00	0.976	0.955
Cal. Specificity	1.00	1.00	1.00
CV Specificity	0.970	0.976	1.000
Cal. Class Er.	0	0	0
Cal. Class Er.	0.01515	0.01219	0.02272

Figure 17: Prediction Ability in **a**: Class 1 **b**: Class 2 and **c**: Class 3 by PLS-DA

3.1.3 Conclusions

Motor oil is a lubricant used in internal combustion engines. Generally, motor oils are composed of base oil (a mixture of one or more base stocks) plus additives to improve the oil's detergency, extreme pressure performance, and ability to inhibit corrosion of engine parts.

Engine oils are evaluated against the American Petroleum Institute requirements (the API sets minimum performance standards for lubricants).

The API categorizes lubricant base stocks into five groups: Groups I-II and III are commonly referred to as mineral oils and group IV is synthetic oil. Group V base stocks are so diverse that there is no catch-all description.

The API service classes [19] have two general classifications: S for "service/spark ignition" and C for "commercial/compression ignition" (typical diesel equipment). Engine oil which has been tested and meets the API standards may display the API Service Symbol (also known as the "Donut") with the service categories on containers sold to oil users. [19] In order to achieve the API mark and certificates specialized laboratories around the world perform some expensive tests to cover all standard requirements for each performance level. [1]

In the present study, NIR and EEM fluorescence spectroscopies have been investigated as alternative solutions to these expensive tests in order to identify the type of base stock in engine oils and to help the formulators when developing a new or tailored lubricant, targeting a given performance level.

PCA performed on the NIR and EEM unfolded spectra showed that base stock samples form clusters according to their API categories and to their chemical composition. PARAFAC outcomes on fluorescence data were in agreement with PCA results.

Partial Least Squares Discriminant Analysis (PLS-DA) applied as a classification technique in order to discriminate the base oil samples according to the first three mineral API categories provided more than satisfactory results: the prediction abilities in the external test set were 87% and 85% using NIR and fluorescence spectroscopy, respectively.

In conclusion, both NIR and fluorescence spectroscopies appeared to be rapid and non-destructive analytical methods for the characterization of base stocks into engine lubricants and they seemed promising tool for Engine Oil Performance Level identification. In particular, NIR spectroscopy proved to be more efficient in the analysis of base stock and motor oils; the reason of this claim is its ability to recognize an increasing and common trend in performance level and API group of base stock existing in the formulation. Furthermore, it was particularly interesting to notice that the position of engine oils projected in the PCs plan computed on the base stock samples was in agreement with the base oil composition in the formulation. In compare with NIR, in engine oil, it appears that the base stock fluorescence spectra are more covered in the presence of additives.

Unfortunately, to collect base stock and motor oil samples was not easy and the classification analysis was affected by this limitation. We therefore intend to continue the study by increasing the number of samples in the different API categories and with different performance levels.

Acknowledgements

We thank Eni S.p.A. (Milan, Italy), Afzoon Ravan Co. (Tehran, Iran) and Bellini SPA (Bergamo, Italy) for providing base stock and engine oil samples.

3.1.5 References

1. Leslie RR (2006) Synthetics, mineral oils, and bio based lubricants. CRC Press Taylor & Francis Group
2. (2011) E-1 rev:01-sep-2011. American Petroleum Institute (API), USA
3. American Petroleum Institute (2012) Engine Oil Licensing and Certification System Seventeenth Edition. 2019
4. (2003) Motor oil. In: Wikimedia. https://en.wikipedia.org/wiki/Motor_oil
5. Wold S, Sjöström M (1972) Statistical Analysis of the Hammett Equation, I. Methods and Model Calculations. Chem Sci 2:49–55
6. Krishnaiah PR (2014) Multivariate Analysis—III: Proceedings of the Third International Symposium on Multivariate Analysis Held at Wright State University, Dayton, Ohio, June 19-24, 1972. Elsevier Science
7. Borin A, Poppi RJ (2004) Multivariate quality control of lubricating oils using Fourier transform infrared spectroscopy. J Braz Chem Soc 15:570–576
8. Amat S, Braham Z, Le Dréau Y, et al (2013) Simulated aging of lubricant oils by chemometric treatment of infrared spectra: Potential antioxidant properties of sulfur structures. Talanta 107:219–224. https://doi.org/10.1016/j.talanta.2012.12.051
9. Hirri A, Bassbasi M, Oussama A (2013) Classification and Quality Control of Lubricating Oils By Infrared Abstract : 3:59–62
10. Leardi R (2009) Experimental design in chemistry: A tutorial. Anal Chim Acta 652:161–172. https://doi.org/10.1016/j.aca.2009.06.015
11. Gareth AL, Didier M, Roger P-T-L (1998) Pharmaceutical Experimental Design. CRC Press Taylor & Francis Group
12. Hooshyari M, Rubio L, Casale M, et al (2019) D-Optimal Design and PARAFAC as Useful Tools for the Optimisation of Signals from Fluorescence

Spectroscopy Prior to the Characterisation of Green Tea Samples. Food Anal Methods 12:761–772. https://doi.org/10.1007/s12161-018-01408-0

13. I. T. Jolliffe (2002) Principal Component Analysis. Springer-Verlag, New York
14. Bro R, Kiers HAL (2003) A new efficient method for determining the number of components in PARAFAC models. J Chemom 17:274–286. https://doi.org/10.1002/cem.801
15. Andersen CM, Mortensen G (2008) Fluorescence Spectroscopy: A Rapid Tool for Analyzing Dairy Products. J Agric Food Chem 56:720–729. https://doi.org/10.1021/jf072025o
16. Murphy KR, Stedmon CA, Graeber D, Bro R (2013) Fluorescence spectroscopy and multi-way techniques. PARAFAC. Anal Methods 5:6557–6566. https://doi.org/10.1039/c3ay41160e
17. Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemom 17:166–173. https://doi.org/10.1002/cem.785
18. Wold S, Ruhe A, Wold H, Dunn WJI (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comput 5:735–743
19. Natick (2019) MATLAB
20. Manson PLS Toolbox
21. Bayes T (1764) An Essay Towards Solving a Problem in the Doctrine of Chances. Int. Encycl. Soc. Sci. . Encycl. 13 Nov. 2019

List of Publications

Published Papers

M. Hooshyari, L. Rubio, M. Casale, S. Furlanetto, F. Turrini, L.A. Sarabia, M.C. Ortiz "D-Optimal Design and PARAFAC as Useful Tools for the Optimisation of Signals from Fluorescence Spectroscopy Prior to the Characterisation of Green Tea Samples", Food Analytical Methods (2019), 12, 761-772.
DOI: https://doi.org/10.1007/s12161-018-01408-0

M. Hooshyari, S. Samanipour, J. A. Baz-Lomba, M. J. Reid, M. Casale, and K. V. Thomas "The Effect of Extraction Methodology on the Recovery and Distribution of Naphthenic Acids of oilfield Produced Water", Science of The Total Environment, (2019), 652, 1416-1423.
DOI: https://doi.org/10.1016/j.scitotenv.2018.10.264

P. Malaspina, M. Casale, C. Malegori, M. Hooshyari, M. Di Carro, E. Magi, P. Giordani "Combining spectroscopic techniques and chemometrics for the interpretation of lichen biomonitoring of air pollution", Chemosphere, (2018), 198, 417-424.
DOI: https://doi.org/10.1016/j.chemosphere.2018.01.136

M. Casale, B. Pasquini, M. Hooshyari, S. Orlandini…" Combining excitation-emission matrix fluorescence spectroscopy, parallel factor analysis, cyclodextrin-modified micellar electrokinetic chromatography and partial least squares class-modelling for green tea characterization", Journal of Pharmaceutical and Biomedical Analysis, (2018), 159, 311-317.
DOI: https://doi.org/10.1016/j.jpba.2018.07.001

Paper to be submitted

Maryam Hooshayari, Paolo Oliveri, Cristina Malegori, Eleonora Mustorgi, Riccardo Leardi, Monica Casale "Identification of Base Stock in Engine Oils by Near Infrared and Fluorescence Spectroscopies coupled with Chemometrics", 2020

List of Presentations

M. Casale, P. Giordani, P. Malaspina, M. Hooshyari, M. Di Carro, “Comparison between NIR spectroscopy and other analytical methods for the bio-monitoring of air pollution by lichens”, ICNIRS 2017, The International Conference on Near Infrared Spectroscopy, Copenhagen, Denmark, 11- 15.06.2017; POSTER.
http://icnirs2017.com

M. Hooshyari, L. Rubio, M. Casale, S. Furnaletto, F. Turrini, L.A. Sarabia, M.C. Ortiz, “D-Optimal Design to Optimize Fluorescent Signals from Solid and Liquid Samples of Green Tea and Their Subsequent Typification” 9th Colloquium Chemiometricum Mediterranean; Arles, France, 27- 30.06.2017; ORAL.
https://colloquim.sciencesconf.org

M. Hooshyari, S. Orlandini, M.C. Ortiz, L.A. Sarabia, S. Furlanetto, “Fluorescence Spectroscopy and Chemometric Techniques for Geographical Discrimination of Green Tea Samples”, XXVI National Conference of the Chemical Italian Society (SCI); Paestum (SA), Italy, 10-14.09.2017; POSTER.
http://sci2017.org

M. Hooshyari, E. Mustorgi, C. Malegori, P. Oliveri, M. Casale. “Effect of storage in plastic bottles on the quality of extra virgin olive oil”, the ninth Giornate Italo-Francesi di Chimica scientific conference, Genova, Italy, 16-18.04 2018; POSTER.
http://gifc2018.sci-liguria.it

M. Casale, M. Hooshyari, E. Mustorgi, C. Malegori, P. Oliveri, “NIR spectroscopy, an efficient tool for evaluating and enhancing the quality of extra virgin olive oil”, VIII Italian Symposium of Spectroscopy NIR, Genova, Italy, 30-31.05.2018; POSTER.
http://www.niritalia2018.sisnir.org

M. Hooshyari*, M. Casale, P. Oliveri, R. Leardi, “Near-Infrared Spectroscopy and Spectrofluorimetry combined with Chemometrics In Order to Determine The Performance Level of Gasoline Engine Oils”, 3rd International Conference and Exhibition on Petroleum, Refining and Environmental Technologies, PEFTEC; Rotterdam, Netherland, 22-23.05,2019; POSTER.
https://www.ilmexhibitions.com/peftec

E. Mustorgi, M. Casale, M. Hooshyari, C. Malegori, P. Oliveri, M. Oteri, L. Mondello, “PLS Regression Models for the Determination of EVOO Quality Parameters by NIR Spectroscopy: a Comparative Study”, 10th Colloquium Chemiometricum Mediterranean, Minorca, Spain 12-14.06.2019; POSTER.
http://www.chemiometricum2019.org

M.Casale, M. Hooshyari, C. Malegori, E. Mustorgi, P. Oliveri, R. Leardi "Spectroscopic Techniques Coupled with Chemometrics for the Identification of Base Oil Type into Engine Oils, 10th Colloquium Chemiometricum Mediterranean, Minorca, Spain 12-14.06.2019; POSTER.
http://www.chemiometricum2019.org

M. Casale, M. Hooshyari, C. Malegori, E. Mustorgi, P. Oliveri, R. Leardi, Comparing Near Infrared Spectroscopy and Spectrofluorimetry in the Determination of Base Oil in Engine Lubricants, 19th International Council for NIR Spectroscopy Meeting, NIR2019, Gold Coast, Australia, 15-20.09.2019; POSTER.
http://www.nir2019.com

E. Mustorgi, M. Casale, M. Hooshyari, P. Oliveri, C. Malegori, R. Bro, S. Furlanetto "Application of PARAFAC on excitation–emission matrix fluorescence spectra for green tea characterisation" XXVIII Congress of the Analytical Chemistry Division Bari 22–26.09.2019; POSTER.
http://barianalitica2019.it

E. Mustorgi, M. Casale, C. Malegori, P. Oliveri, M. Hooshyari, L. Mondello, M. Oteri ”Evaluation of analytical performances of quartz cuvettes and disposable glass vials for the determination of fame and tags in extra virgin olive oil." XXVIII Congress of the Analytical Chemistry Division Bari 22 – 26.09.2019; POSTER.
http://barianalitica2019.it

M. Casale, M. Hooshyari, C. Malegori, E. Mustorgi, P. Oliveri "Characterization of base oils for engine lubricants by nir and fluorescence spectroscopies coupled with chemometrics" XXVIII Congress of the Analytical Chemistry Division Bari 22 – 26.09.2019; ORAL.
http://barianalitica2019.it

Appendix 1

List of the 63 GT samples analysed by EEM fluorescence spectroscopy.

Sample Code[i]	Name	Type/Zone[ii]
J1	Bancha	Bancha
J2	Gyokuro	Gyokuro
J3	Houjicha	Bancha Bio
J4	Matcha Tsuru	Matcha Tsuru
J5	Matcha Tsuru	Matcha Tsuru
J6	Bancha	Bancha
J7	Sencha	Sencha
J8	Matcha	Matcha
J9	Kukicha	Sencha
J10	Kukicha	Sencha
J11	Kukicha	Sencha
J12	Bancha	Bancha
J13	Bancha	Bancha
J14	Bancha	Bancha
J15	Sencha	Sencha
J16	Sencha-Matcha	Sencha/Matcha
J17	Bancha-Hojicha	Bancha
J18	Matcha	Matcha
J19	Sencha-Matcha	Sencha/Matcha
J20	Sencha	Sencha
J21	Bancha	Bancha
J22	Matcha	Matcha
J23	Kokeicha Green	Matcha
J24	Tamariokucha	Sencha
J25	Matcha Tsuki	Matcha
J26	Matcha Tsuki	Matcha
J27	Matcha Kotobuki	Matcha
J28	Matcha Kotobuki	Matcha
J29	Sencha Special Fine	Sencha
C1	King Jasmine	Hunan
C2	Jasmine Dragon	Fujian
C3	Snow Bud	Zhejiang
C4	Snow Bud	Zhejiang
C5	Gunpowder	Zhejiang
C6	Mistery Rose	Fujian
C7	Mistery Rose	Fujian
C8	Mao Feng	Anhui
C9	Yunnan Green	Yunnan
C10	Yunnan Green	Yunnan
C11	White Monkey Pekoe	Fujian
C12	China Li Zi Yang	Guandong
C13	Gu Zhan Mao Jian	Hunan
C14	Pi Lo Chun	Jiangsu
C15	Palace Needle	Hubei
C16	Mini Tuo Cha	Yunnan
C17	Mini Tuo Cha	Yunnan
C18	White Heart	Fujian
C19	White Heart	Fujian
C20	Tai Mu Long Zhu	Fujian
C21	Lung Ching Top Grade	Zhejiang
C22	Yellow Sunshine	Shandong
C23	Special Gunpowder	Zhejiang
C24	Lung Ching Special	Zhejiang
C25	Dong Yang Dong Bai	Zhejiang
C26	Green Tea OP	Fujian
C27	Jasmine Special	Fujian
C28	Xia Zhou Bi Feng	Hubei
C29	Special Gunpowder Tea	Zhejiang
C30	Green Magnolia	Jiangsu
C31	Sweet Osmanto	Guanxi
C32	Yong Xi Hou Quing	Anhui
C33	Jasmine Chung Feng	Fujian
C34	Silver Sprout Green	Hunan

[i] J: Japanese GT samples; C: Chinese GT samples.

[ii] Different types are reported for Japanese GT samples, and different geographical zones are reported for Chinese GT samples.

Appendix 2

List of EVOO samples analysed.

No. of Sample	Company	Year	Olives	Origin
1	SABINO LEONE	2017/2018	CORATINA-REGINA DELLA PUGLIA monocultivar	APULIA
2	AZIENDA AGRICOLA COSTANTINO MARIELLA	2017	PERANZANA monovarietal	APULIA
3	SOCIETA' AGRICOLA DEMAR S.R.L.	2017/2018	CORATINA monocultivar	APULIA
4	SABINO LEONE	2017/2018	FRANTOIO monocultivar	APULIA
5	ELAIOPOLIO COOP RIFORMA FONDIARIA SCA	2017/2018	CORATINA monocultivar	APULIA
6	ELAIOPOLIO COOP RIFORMA FONDIARIA SCA	2017/2018	PERANZANA monocultivar	APULIA
7	AZIENDA AGRICOLA DE CARLO SOCIETA' AGRICOLA SEMPLICE	2017/2018	OGLIAROLA (CIMA DI BITONTO) 100%	APULIA
8	SCIROCCO AZIENDA AGRICOLA	2017/2018	CERASUOLA - NOCELLARA DEL BELICE - BIANCOLILLA	SICILY
9	ANTONINO CENTONZE	2017	MONOCULTIVAR NOCELLARA DEL BELICE	SICILY
10	FRANTOIO CUTRERA	2017/2018	MONOCULTIVAR TONDA IBLEA 100%	SICILY
11	AZIENDA AGRICOLA FATTORIA SANT'ANASTASIA	2017/2018	NOCELLARA MESSINESE MONOCULTIVAR	SICILY
12	AZIENDA AGRICOLA FATTORIA SANT'ANASTASIA	2017/2018	NOCELLARA ETNA MONOCULTIVAR	SICILY
13	ROMANO VINCENZO & C. SAS	2017/2018	NOCELLARA ETNEA	SICILY
14	AZIENDA AGRICOLA TORNATURI CARMELA	2017	NOCELLARA DEL BELICE MONOCULTIVAR	SICILY
15	TENUTA GALLINELLA DI PIETRO SABELLA	2017/2018	BIANCOLILLA	SICILY
16	DIEVOLE SRL	2017	LECCINO-MORAIOLO-FRANTOIO-MAURINO	TUSCANY
17	AZIENDA AGRARIA GIANCARLO GIANNINI	2017	MORAIOLO-FRANTOIO-LECCINO	TUSCANY
18	LOGGIA DEL CENTONE	2017/2018	FRANTONIO E LECCINO	TUSCANY
19	AGR. POTASSA SRL	2017	MORAIOLO, LECCINO, FRANTOIO, CORREGGIOLO, OLIVASTRA	TUSCANY
20	OLIVIERA SANT'ANDREA DI GIGANTI E & E SNC	2017	CORREGGIOLO 50%, PENDOLINO 10%, MAURINO 20%, LECCIO DEL CORNO 20%	TUSCANY

No. of Sample	Company	Year	Olives	Origin
21	IL CORNO S.A.R.L	2017	MORAIOLO, LECCINO, FRANTOIO, altre cultivar minori	TUSCANY
22	FRANTOIO FRANCI SNC	2017	FRANTOIO MONOCULTIVAR	TUSCANY
23	AZIENDA AGRICOLA IL TORRIANO SNC	2017	FRANTOIO 40% - MORAIOLO 35% - LECCINO 20%- PENDOLINO 5%	TUSCANY
24	FATTORIA CASTEL RUGGERO	2017	FRANTOIO - LECCINO E MORAIOLO insieme oltre l'80% + PENDOLINO, MORCHIAIO, LECCIO del Corno	TUSCANY
25	AZIENDA POGGIO TORSELLI SRL SOCIETA' AGRICOLA GALLARATE	2017	FRANTOIO MORAIOLO LECCINO PENDOLINO	TUSCANY
26	AZIENDA AGRICOLA LOSI PONTIGLIANELLO	2017	CORREGGIOLO 95%- LECCINO 5%- FRANTOIANO 5%	TUSCANY
27	SOCIETA' AGRICOLA DI FOIANO DI GAETANO PAOLO E SIMONE SS	2017/2018	FRANTOIO 100%	TUSCANY
28	AZIENDA AGRICOLA LA COSTA S.S.A.	2017	MORAIOLO monovarietal	TUSCANY
29	IL FELCIAIO SSA DI FERRINI SANDRO E LUIGI	2017/2018	FRANTOIO monocultivar	TUSCANY
30	FRANTOIO FRANCI SNC	2017	FRANTOIO monocultivar	TUSCANY
31	OLIO DI DIEVOLE SRL	2017/2018	CORATINA MONOCULTIVAR	TUSCANY
32	PODERE SANTA GIULIA	2017	LECCIO DEL CORNO 100%	TUSCANY
33	AZIENDA AGRICOLA SOLAIA DI BROGELLI e C.	2017	LECCIO DEL CORNO monovarietal	TUSCANY
34	L'ANTICO FRANTOIO DI SEGALARI	2017	LECCIO DEL CORNO monovaietale	TUSCANY
35	OLIO DI DIEVOLE SRL	2017/2018	NOCELLARA monocultivar	TUSCANY
36	SOCIETA' AGRICOLA FELSINA SpA	2017	RAGGIOLO MONOCULTIVAR	TUSCANY
37	FATTORIA RAMERINO SOCIETA' AGRICOLA	2017	MORAIOLO	TUSCANY
38	FATTORIA ALTOMENA SRL	2017	FRANTOIO monocultivar	TUSCANY
39	FATTORIA CORZANO E PATERNO	2017	FRANTOIO monocultivar	TUSCANY
40	FATTORIA CORZANO E PATERNO	2017	PENDOLINO monocultivar	TUSCANY
41	TENUTA DI ARTIMINO SOCIETA' AGRICOLA SRL	2017/2018	FRANTOIO, LECCINO, MORAIOLO, altre varieta' minori	TUSCANY
42	FRANTOIO DEL GREVEPESA	2017	FRANTOIO(principale 50-60%), LECCINO, MORAIOLO, PENDOLINO piccole quantita'	TUSCANY
43	PODERE SANTA GIULIA	2017/2018	NON INDICATA (info@ilcavallino.it)	TUSCANY

No. of Sample	Company	Year	Olives	Origin
44	PAOLO CASSINI	2017/2018	TAGGIASCA MONOCULTIVAR	LIGURIA
45	IL CAVALIERE	2017	CASALIVA cultivar principale, LECCINO, MORAIOLO, PENDOLINO, FRANTOIO	LOMBARD Y
46	AGRARIA RIVA DEL GARDA FRANTOIO DI RIVA	2017	CASALIVA monovarietal	TRENTINO
47	AGRARIA RIVA DEL GARDA FRANTOIO DI RIVA	2017	CASALIVA (>70%), FRANTOIO LECCINO (2-3%)	TRENTINO
48	SOCIETA' AGRICOLA TENUTA POJANA	2017/2018	GRIGNANO, FAVAROL, PENDOLINO e TREPP	VENETO
49	SOCIETA' AGRICOLA TENUTA POJANA	2017/2018	7 DIFFERENTI CULTIVAR	VENETO
50	AZIENDA AGRICOLA CONFORTI GIUSEPPE	2017/2018	NON INDICATA (info@agricolaconforti.it)	CALABRIA
51	FRANCESCA DE LEO ALBERTI	2017/2018	OTTOBRATICA, SINOPOLESE	CALABRIA
52	OLEARIA S. GIORGIO F.LLI FAZARI S.N.S	2017/2018	OTTOBRATICA MONOCULTIVAR	CALABRIA
53	AZIENDA AGRICOLA SORELLE GARZO	2017	OTTOBRATICA MONOCULTIVAR	CALABRIA
54	SANTA TECLA AZIENDA AGRICOLA DI RITA LICASTRO	2017/2018	OTTOBRATICA monovarietal	CALABRIA
55	AZIENDA AGRICOLA SORELLE GARZO	2017	OTTOBRATICA, SINOPOLESE	CALABRIA
56	VILLA CAVICIANA SS	2017/2018	CANINESE 100%	LATIUM
57	SOCIETA' AGRICOLA COLLI ETRUSCHI	2017	CANINESE monovarietal	LATIUM
58	SOCIETA' AGRICOLA COLLI ETRUSCHI	2017	CANINESE, FRANTOIO, MAURINO	LATIUM
59	FRANTOIO TUSCUS DI GIAMPAOLO SODANO e C. SAS	2017	Leccino/bolzone	LATIUM
60	FRANTOIO TUSCUS DI GIAMPAOLO SODANO e C. SAS	2017	NON INDICATA (info@frantoiotuscus.com)	LATIUM
61	SANTINA DELLE FATE SOC. COOP	2017/2018	ITRANA monovarietal	LATIUM
62	ACCADEMIA OLEARIA SRL	2017/2018	BOSANA, SEMIDANA	SARDINIA
63	ACCADEMIA OLEARIA SRL	2017/2018	BOSANA in prevalenza	SARDINIA
64	ACCADEMIA OLEARIA SRL	2017/2018	BOSANA monovarietal	SARDINIA
65	AZIENDA AGRICOLA CANNAVERA	2017/2018	BOSANA monovarietal	SARDINIA
66	AZIENDA AGRICOLA EUGENIO RANCHINO	2017	LECCINO 60%-FRANTOIO 30%-MORAIOLO NON SUPERIORE AL 15%	UMBRIA
67	AZIENDA AGRICOLA GIULIO MANNELLI	2017/2018	MORAIOLO-FRANTOIO-LECCINO	UMBRIA
68	AZIENDA AGRICOLA ADRIATICA VIVAI	2017/2018	CORATINA monocultivar	APULIA

No. of Sample	Company	Year	Olives	Origin
69	FRANTOIO DI BINETTO (BA) SP BITETTO-BINETTO (BA) da SCHIRALLI SRL	2017/2018	CORATINE monocultivar	APULIA
70	AZIENDA AGRICOLA DEPALO LUIGI	2017/2018	CORATINA 100%	APULIA
71	AZIENDA AGRICOLA DEPALO LUIGI	2017/2018	OGLIAROLA 100%	APULIA
72	SCHIRALLI SRL NEL FRANTOIO DI BINETTO (BA) S.P. BITETTO-BINETTO	2017/2018	OGLIAROLA	APULIA
73	AZIENDA AGRICOLA DONATO CONSERVA	2017/2018	PARANZANA monocultivar	APULIA
74	AZIENDE AGRICOLE PLANETA SS	2017	NOCELLARA DEL BELICE-BIANCOLILLA-CERASUOLA	SICILY
75	SOCIETA' AGRICOLA VERNERA DI SPANO' & C. SNC	2017/2018	MONOCULTIVAR TONDA IBLEA	SICILY
76	PRUNETI	2017/2018	LECCINO-MORAIOLO-FRANTOIO, varietà minori	TUSCANY
77	CALDINI GUIDO SRL PODERE DI VENTURINA	2017/2018	FRANTOIO 40%-MORAIOLO 30%-LECCINO 30%	TUSCANY
78	CIACCI ANNA PODERE VIGNINE	2017	OLIVASTRA SAGGIANESE (autoctona)	TUSCANY
79	ADMEATA DI JEAN CLAUDE ZACCHINI	2017	OLIVASTRA SAGGIANESE MONOCULTIVAR	TUSCANY
80	SOCIETA' AGRICOLA LA CROCETTA	2017	FRANTOIO - LECCINO - MORAIOLO - PENDOLINO	TUSCANY
81	LE CORTI S.p.A	2017	FRANTOIO cultivar prevalente, MORAIOLO, LECCINO	TUSCANY
82	MARCHESI MAZZEI SPA AGRICOLA	2017	MORAIOLO- 50% LECCINO- 50%	TUSCANY
83	AZIENDE BARONE RICASOLI SPA AGRICOLA	2017	FRANTOIO monovarietal	TUSCANY
84	PODERE GIACOMO GRASSI	2017	FRANTOIO monovarietal	TUSCANY
85	PODERE GIACOMO GRASSI	2017	PENDOLINO monovarietal	TUSCANY
86	OLIVIERO TOSCANI SOCIETA' AGRICOLA SRL	2017	MORAIOLO monocultivar	TUSCANY
87	OLIVART	2017	MORAIOLO monovarietal	TUSCANY
88	OLIVART	2017	LECCINO monovarietal	TUSCANY
89	OLIVART	2017	FRANTOIO monovarietal	TUSCANY
90	SOCIETA' AGRICOLA PODERE VAL D'ORCIA SRL	2017/2018	MAURINO monovarietal	TUSCANY
91	SOCIETA' AGRICOLA FELSINA SpA	2017	PENDOLINO MONOCULTIVAR	TUSCANY
92	BARALDI DIEGO	2017	CASALIVA monovarietal	LOMBARD Y
93	OLIOCRU SRL	2017/2018	CASALIVA MONOCULTIVAR	TRENTINO
94	MONTENIGO	2017/2018	GRIGNANO monovarietal	VENETO
95	NICOTERA SEVERISIO FERDINANDO SS AGRICOLA	2017/2018	CAROLEA 100%	CALABRIA

No. of Sample	**Company**	**Year**	**Olives**	**Origin**
96	COSMO DI RUSSO - Via Pontone snc 04024 Gaeta (LT)	2017/2018	ITRANA monovarietal	LATIUM
97	IL MOLINO SOCIETA' AGRICOLA SCIUGA SS - Via del Lago km 5 01027 Montefascone (VT)	2017/2018	CANINO (DENOCCIOLATO)	LATIUM
98	QUATTROCIOCCHI AMERICO - Via Mole Santa MaRIA 03011 Alatri (FR)	2017/2018	ITRANA 100%	LATIUM
99	IL MOLINO SOCIETA' AGRICOLA SCIUGA SS - Via del Lago km 5 01027 Montefascone (VT)	2017/2018	CANINO monovarietal	LATIUM
100	IL MOLINO SOCIETA' AGRICOLA SCIUGA SS - Via del Lago km 5 01027 Montefascone (VT)	2017/2018	FRANTOIO monovarietal	LATIUM
101	IONE ZOBI SRL	2017	CANINESE monovarietal	LATIUM
102	AZIENDA AGRICOLA SEBASTIANO FADDA	2018	NERA DI OLIENA MONOCULTIVAR	SARDINIA
103	AZIENDA AGRICOLA CANNAVERA	2017/2018	BOSANA MONOCULTIVAR	SARDINIA
104	MONINI SPA SS	2017/2018	CORATINA MONOCLTIVAR	UMBRIA
105	MONINI SPA SS	2017/2018	FRANTOIO MONOCLTIVAR	UMBRIA
106	OLIO METELLI SAS	2017	MORAIOLO MONOCULTIVAR	UMBRIA

Appendix 3

Characterization of naphthenic acids in oil sands wastewaters by gas chromatography-mass spectrometry.

Formula (CnH2n+zO2)	n	z	Cycle	Mass –theoretical Kmass (Da)	Mass-H
C8H14O2	8	-2	1	142.1164	141.0918
C8H16O2	8	0	0	144.1164	143.1074
C9H14O2	9	-4	2	154.1322	153.0918
C9H16O2	9	-2	1	156.1322	155.1074
C9H18O2	9	0	0	158.1322	157.1231
C10H16O2	10	-4	2	168.148	167.1074
C10H18O2	10	-2	1	170.148	169.1231
C10H20O2	10	0	0	172.148	171.1387
C11H16O2	11	-6	3	180.1638	179.1074
C11H18O2	11	-4	2	182.1638	181.1231
C11H20O2	11	-2	1	184.1638	183.1387
C11H22O2	11	0	0	186.1638	185.1544
C12H18O2	12	-6	3	194.1796	193.1231
C12H20O2	12	-4	2	196.1796	195.1387
C12H22O2	12	-2	1	198.1796	197.1544
C12H24O2	12	0	0	200.1796	199.17
C13H18O2	13	-8	4	206.1954	205.1231
C13H20O2	13	-6	3	208.1954	207.1387
C13H22O2	13	-4	2	210.1954	209.1544
C13H24O2	13	-2	1	212.1954	211.17
C13H26O2	13	0	0	214.1954	213.1857
C14H20O2	14	-8	4	220.2112	219.1387
C14H22O2	14	-6	3	222.2112	221.1544
C14H24O2	14	-4	2	224.2112	223.17
C14H26O2	14	-2	1	226.2112	225.1857
C14H28O2	14	0	0	228.2112	227.2014
C15H20O2	15	-10	5	232.227	231.1387
C15H22O2	15	-8	4	234.227	233.1544
C15H24O2	15	-6	3	236.227	235.17
C15H26O2	15	-4	2	238.227	237.1857
C15H28O2	15	-2	1	240.227	239.2014
C15H30O2	15	0	0	242.227	241.217
C17H22O2	16	-10	5	246.2428	245.1544
C17H24O2	16	-8	4	248.2428	247.17
C16H26O2	16	-6	3	250.2428	249.1857
C16H28O2	16	-4	2	252.2428	251.2014
C16H30O2	16	-2	1	254.2428	253.217
C16H32O2	16	0	0	256.2428	255.2327
C17H22O2	17	-12	6	258.2586	257.1544
C17H24O2	17	-10	5	260.2586	259.17
C17H26O2	17	-8	4	262.2586	261.1857
C17H28O2	17	-6	3	264.2586	263.2014
C17H30O2	17	-4	2	266.2586	265.217
C17H32O2	17	-2	1	268.2586	267.2327
C17H34O2	17	0	0	270.2586	269.2483
C18H24O2	18	-12	6	272.2744	271.17

Formula (CnH2n+zO2)	**n**	**z**	**Cycle**	**Mass –theoretical Kmass (Da)**	**Mass-H**
C18H26O2	18	-10	5	274.2744	273.1857
C18H28O2	18	-8	4	276.2744	275.2014
C18H30O2	18	-6	3	278.2744	277.217
C18H32O2	18	-4	2	280.2744	279.2327
C18H34O2	18	-2	1	282.2744	281.2483
C18H36O2	18	0	0	284.2744	283.264
C19H26O2	19	-12	6	286.2902	285.1857
C19H28O2	19	-10	5	288.2902	287.2014
C19H30O2	19	-8	4	290.2902	289.217
C19H32O2	19	-6	3	292.2902	291.2327
C19H34O2	19	-4	2	294.2902	293.2483
C19H36O2	19	-2	1	296.2902	295.264
C19H38O2	19	0	0	298.2902	297.2796
C20H28O2	20	-12	6	300.306	299.2014
C20H30O2	20	-10	5	302.306	301.217
C20H32O2	20	-8	4	304.306	303.2327
C20H34O2	20	-6	3	306.306	305.2483
C20H36O2	20	-4	2	308.306	307.264
C20H38O2	20	-2	1	310.306	309.2796
C20H40O2	20	0	0	312.306	311.2953
C21H32O2	21	-12	6	314.3218	313.217
C21H32O2	21	-10	5	316.3218	315.2327
C21H34O2	21	-8	4	318.3218	317.2483
C21H36O2	21	-6	3	320.3218	319.264
C21H38O2	21	-4	2	322.3218	321.2796
C21H240O2	21	-2	1	324.3218	323.2953
C21H42O2	21	0	0	326.3218	325.3109
C22H30O2	22	-12	6	328.3376	327.2327
C22H32O2	22	-10	5	330.3376	329.2483
C22H34O2	22	-8	4	332.3376	331.264
C22H38O2	22	-6	3	334.3376	333.2796
C22H40O2	22	-4	2	336.3376	335.2953
C22H42O2	22	-2	1	338.3376	337.3109
C22H44O2	22	0	0	340.3376	339.3266
C23H34O2	23	-12	6	342.3534	341.2483
C23H36O2	23	-10	5	344.3534	343.264
C23H38O2	23	-8	4	346.3534	345.2796
C23H40O2	23	-6	3	348.3534	347.2953
C23H42O2	23	-4	2	350.3534	349.3109
C23H44O2	23	-2	1	352.3534	351.3266
C23H46O2	23	0	0	354.3534	353.3423
C24H36O2	24	-12	6	356.3692	355.264
C24H38O2	24	-10	5	358.3692	357.2796
C24H40O2	24	-8	4	360.3692	359.2953
C24H42O2	24	-6	3	362.3692	361.3109
C24H44O2	24	-4	2	364.3692	363.3266
C24H46O2	24	-2	1	366.3692	365.3423
C24H48O2	24	0	0	368.3692	367.3579
C25H38O2	25	-12	6	370.385	369.2796
C25H40O2	25	-10	5	372.385	371.2953
C25H42O2	25	-8	4	374.385	373.3109
C25H44O2	25	-6	3	376.385	375.3266
C25H46O2	25	-4	2	378.385	377.3423
C25H48O2	25	-2	1	380.385	379.3579
C25H50O2	25	0	0	382.385	381.3736
C26H40O2	26	-12	6	384.4008	383.2953

Formula (CnH2n+zO2)	n	z	Cycle	Mass –theoretical Kmass (Da)	Mass-H
C26H42O2	26	-10	5	386.4008	385.3109
C26H44O2	26	-8	4	388.4008	387.3266
C26H46O2	26	-6	3	390.4008	389.3423
C26H48O2	26	-4	2	392.4008	391.3579
C26H50O2	26	-2	1	394.4008	393.3736
C26H52O2	26	0	0	396.4008	395.3892
C27H42O2	27	-12	6	398.4166	397.3109
C27H44O2	27	-10	5	400.4166	399.3266
C27H46O2	27	-8	4	402.4166	401.3423
C27H48O2	27	-6	3	404.4166	403.3579
C27H50O2	27	-4	2	406.4166	405.3736
C27H52O2	27	-2	1	408.4166	407.3892
C27H54O2	27	0	0	410.4166	409.4049
C28H44O2	28	-12	6	412.4324	411.3266
C28H46O2	28	-10	5	414.4324	413.3423
C28H48O2	28	-8	4	416.4324	415.3579
C28H50O2	28	-6	3	418.4324	417.3736
C28H52O2	28	-4	2	420.4324	419.3892
C28H54O2	28	-2	1	422.4324	421.4049
C28H56O2	28	0	0	424.4324	423.4205
C29H46O2	29	-12	6	426.4482	425.3423
C29H48O2	29	-10	5	428.4482	427.3579
C29H50O2	29	-8	4	430.4482	429.3736
C29H52O2	29	-6	3	432.4482	431.3892
C29H54O2	29	-4	2	434.4482	433.4049
C29H56O2	29	-2	1	436.4482	435.4205
C29H58O2	29	0	0	438.4482	437.4362
C30H48O2	30	-12	6	440.464	439.3579
C30H50O2	30	-10	5	442.464	441.3736
C30H52O2	30	-8	4	444.464	443.3892
C30H54O2	30	-6	3	446.464	445.4049
C30H56O2	30	-4	2	448.464	447.4205
C30H58O2	30	-2	1	450.464	449.4362
C30H60O2	30	0	0	452.464	451.4519
C31H50O2	31	-12	6	454.4798	453.3736
C31H52O2	31	-10	5	456.4798	455.3892
C31H54O2	31	-8	4	458.4798	457.4049
C31H56O2	31	-6	3	460.4798	459.4205
C31H58O2	31	-4	2	462.4798	461.4362
C31H60O2	31	-2	1	464.4798	463.4519
C31H62O2	31	0	0	466.4798	465.4675
C32H52O2	32	-12	6	468.4956	467.3892
C32H54O2	32	-10	5	470.4956	469.4049
C32H56O2	32	-8	4	472.4956	471.4205
C32H58O2	32	-6	3	474.4956	473.4362
C32H60O2	32	-4	2	476.4956	475.4519
C32H62O2	32	-2	1	478.4956	477.4675
C32H64O2	32	0	0	480.4956	479.4832
C33H54O2	33	-12	6	482.5114	481.4049
C33H56O2	33	-10	5	484.5114	483.4205
C33H58O2	33	-8	4	486.5114	485.4362
C33H60O2	33	-6	3	488.5114	487.4519
C33H62O2	33	-4	2	490.5114	489.4675
C33H64O2	33	-2	1	492.5114	491.4832
C33H66O2	33	0	0	494.5114	493.4988
C34H56O2	34	-12	6	496.5272	495.4205

Formula (CnH2n+zO2)	n	z	Cycle	Mass –theoretical Kmass (Da)	Mass-H
C34H58O2	34	-10	5	498.5272	497.4362
C34H60O2	34	-8	4	500.5272	499.4519
C34H62O2	34	-6	3	502.5272	501.4675
C34H64O2	34	-4	2	504.5272	503.4832
C34H66O2	34	-2	1	506.5272	505.4988
C34H68O2	34	0	0	508.5272	507.5145
C35H58O2	35	-12	6	510.543	509.4362
C35H60O2	35	-10	5	512.543	511.4519
C35H62O2	35	-8	4	514.543	513.4675
C35H64O2	35	-6	3	516.543	515.4832
C35H66O2	35	-4	2	518.543	517.4988
C35H68O2	35	-2	1	520.543	519.5145
C35H70O2	35	0	0	522.543	521.5301
C23H39O3	23	-6	3	362.8847	361.8768
C22H35O4	22	-8	4	362.8485	361.8406
C22H35O2S	22	-8	4	362.8307	361.8228
C25H31O2	25	-18	9	362.8341	361.8262
C25H31S	25	-18	9	362.8095	361.8016
C21H31O3S	21	-10	5	362.7943	361.7864
C24H27OS	24	-20	10	362.7733	361.7654
C20H27O4S	20	-12	6	362.758	361.7501
C23H23O2S	23	-22	11	362.7369	361.729
C19H23O3S2	19	-14	7	362.7039	361.696

Appendix 4

Base Stock Samples.

No.	API Group	Experiment Code	No.	API Group	Experiment Code
1	Group I – Solvent refined	150Brte	28	Group III	BO11a
2		SN150a	29		BO12e
3		SN150e	30		BO13a
4		SN500e	31		BO14e
5		SN01b	32		BO15e
6		SN02b	33		BO16e
7		SN03b	34		BO17e
8		SN04b	35		BO18e
9		SN05b	36		BO19e
10		SN06b	37		BO18b
11		SN07b	38		BO19b
12		SN08b	39		BO20b
13		SN09b	40		BO21b
14		SN10b	41		BO22b
15		SN11b	42		BO23b
16		SN12b	43		BO24b
17		SN13b	44		BO25
18		SN14b	45		BO26b
19		SN15b	46		BO27b
20	Group II	500Na	47		BO28b
21		150Ne	48	Group IV	BO20e
22		400Ne	49		BO21e
23		300Ne	50		BO22e
24		180Na	51		BO29b
25		350Na	52	Group I Re-refined	RBO23e
26		16Nb	53		RBO24e
27		17Nb			

Appendix 5

Engine Oil Samples and Their Base Oil Composition.

No.	**Experiment Code**	**Group of Base Stock**				
	(Performance Level)	**Group I**	**Group II**	**Group III**	**Group IV**	**Re-refined**
1	**MIX 1033**	100%				
2	**MIX 1084**					100%
3	**MIX 1035**			38.7%		61.3%
4	**MIX 1022**			100%		
5	**MIX 1016**			70.5%	29.5%	
6	**MIX 1024**		55.6%	43.1%	1.3%	
7	**MIX 1302**	84%			16%	
8	**MIX 1042**	96.5%		3.5%		
9	**MIX 1086**	79.8%		20.2%		
10	**MIX1101** (SL)	100%				
11	**MIX1102** (SL)		100%			
12	**MIX1103** (SN)			30.05%	69.95%	
13	**MIX1104** (SN)			100%		
14	**MIX1105** (SG/CD)		100%			
15	**MIX1106** (SJ/CF)		100%			
16	**MIX1107** (SM)			100%		
17	**MIX1108** (SG/CD)		100%			
18	**MIX1109** (SM)		100%			
19	**MIX1110** (SM)			100%		
20	**MIX1201** (CH4)		100%			
21	**MIX1202** (CI4)			100%		
23	**MIX1203** (CF)		100%			
24	**MIX1204** (CI4)		87.21%	12.79%		
25	**MIX1205** (CI4)	100%				
26	**MIX1206** (CF)	29.59%	70.41%			
27	**MIX1430** (SL)	☼		☼		
28	**MIX1431** (SL)	☼		☼		
29	**MIX1432** (SJ)			100%		
30	**MIX1433** (SG)	100%				
31	**MIX1434** (SG)	100%				
32	**MIX1435** (SG)	100%				
33	**MIX1436** (SG)	100%				
34	**MIX1437** (SL)	☼		☼		
35	**MIX1438** (SJ)			100%		
36	**MIX1439** (SJ)			100%		
37	**MIX1440** (SJ)			100%		
38	**MIX1441** (SL)			100%		
39	**MIX1442** (SG)			100%		
40	**MIX1443**(SN)				100%	
41	**MIX1444** (SJ)			100%		
42	**MIX1445** (SJ)			100%		
43	**MIX1446** (SA)		100%			

☼: The treat rate is protected by the producer.