TÉCNICO LISBOA

Identification and quantification of unwanted coloring agents through Spectroscopy and Chemometrics

Sara Ricardo Carvalho Mesquita

Thesis to obtain the Master of Science Degree in

Biological Engineering

Supervisors: Prof. José Monteiro Cardoso de Menezes and Prof. Åsmund Rinnan

Examination Committee

Chairperson: Prof. Arsénio do Carmo Sales Mendes Fialho

Supervisor: Prof. José Monteiro Cardoso de Menezes

Member of the Committee: Prof. Marília Clemente Velez Mateus

December 2014

ACKNOWLEDGEMENTS

A very special thanks to my supervisor at the Faculty of Life Science, University of Copenhagen, Associated Professor Åsmund Rinnan, for providing me an excellent guidance and support throughout the project with all his knowledge, kindness and encouragement.

Many thanks to everyone from the SPECC group at the University of Copenhagen for being so friendly and always willing to help.

Furthermore, I would also like to thank my supervisors at Chr. Hansen A/S, Department manager at Color R&D Morten Eriksen and Research scientist Jens Møller for providing me helpful information for my thesis and for the great visit to the Chr.Hansen production plant.

Additionally, I would like to offer my sincerest gratitude to my supervisor at IST, Professor José Cardoso Menezes, for raising my interest in the area of Chemometrics, which led to my involvement in this project at SPECC.

Finally, a special thanks to my family and friends, namely Sofia Santos and Ricardo Caroço, for the great moral support during this project.

Sara Mesquita

RESUMO

O objectivo deste projecto é o desenvolvimento de um modelo rápido e não invasivo para detectar e quantificar corantes indesejados - *Orange II*, *Tartrazine*, *Sunset Yellow* e *Annatto* – no corante alimentar amarelo extraído de *Carthamus tinctorius* L. (açafrão-bastardo).

Para tal, três métodos de espectroscopia – Ultravioleta-Vísivel (UV-Vis), Fluorescência e Infravermelho próximo (NIR) – foram investigados quanto à sua utilidade na detectação da presença dos aditivos indesejados em *Carthamus* em diferentes concentrações. Posteriormente, realizou-se análise multivariada aos dados adquiridos com o objectivo de criar os modelos pretendidos.

Primeiramente, a análise exploratória de dados, utilizando *Principal Component Analysis* (PCA), foi executada para os três métodos espectroscópicos. Seguidamente, os dados que mostraram resultados satisfatórios no PCA – dados do UV-Visível e da Fluorescência - foram utilizados para implementar *Partial Least Squares (PLS) Regression* com o objectivo de criar um modelo quantitativo. Por fim, realizou-se uma análise qualitativa utilizando os mesmo dados utilizados no PLS. Para o UV-Vísivel, os métodos de classificação utilizados foram *PLS–Discriminant Analysis* (PLS-DA) e *Soft Independent Modeling by Class Analogy* (SIMCA) e para a Fluorescência, PLS-DA e *k-Nearest Neighbors (k*-NN*)* foram implementados.

Verificou-se que não é possível criar um modelo quantitativo eficiente devido ao limite de detecção das baixas concentrações utilizadas. Para classificação, os dados do UV-Vísivel mostraram resultados superiores. No entanto, ao analisar os espectros da Fluorescencência, observou-se que um modelo mais preciso poderia ser criado se um conjunto de dados mais abrangente fosse utilizado. Em investigações futuras, mais amostras puras e impuras deverão ser medidas utilizando um método de Fluorescência optimizado.

PALAVRAS-CHAVE *Carthamus tinctorius*, Espectroscopia, Utravioleta-Visível, Fluorescência, Quimiometria, PLS-DA, SIMCA, *k*-NN.

ABSTRACT

This project focus on developing a fast and non-invasive method able to detect and quantify the presence of unwanted coloring additives – Orange II, Tartrazine, Sunset Yellow and Annatto - in a yellow food colorant extracted from *Carthamus tinctorius* L. (Safflower).

Therefore, three spectroscopic methods – Ultraviolet-Visible (UV-Vis), Fluorescence and Near-Infrared (NIR) – were investigated for their usefulness to detect the presence of undesired additives in *Carthamus* in different concentrations. Subsequently, chemometric tools were implemented on the data acquired with the aim of creating the necessary models.

Firstly, exploratory data analysis, using Principal Component Analysis (PCA), was executed on the data acquired with the three spectroscopic methods. Thereafter, the data that showed a good behavior in PCA - UV-Visible and Fluorescence data - was used to perform Partial Least Squares (PLS) Regression in order to create a quantitative model. Finally, a qualitative analysis was made for the same data used in PLS. The classification methods used for UV-Visible were PLS-Discriminant Analysis (PLS-DA) and Soft Independent Modeling by Class Analogy (SIMCA) and for Fluorescence, PLS-DA and *k*-Nearest Neighbors (*k*-NN) were implemented.

It was verified that is not possible to create a valuable quantitative model, due to the detection limitations concerning the low concentrations used. For classification purposes, the UV-Visible data showed superior results. However, by analyzing the Fluorescence spectra, it was observed that a more accurate classification model could be created if a more comprehensive data set was used. In future work, more pure and impure samples should be measured in an optimized Fluorescence method.

KEYWORDS *Carthamus tinctorius*, Spectroscopy, Ultraviolet-Visible, Fluorescence, Chemometrics, Classification, PLS-DA, SIMCA, *k*-NN.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADI	Acceptable Daily Intake
CV	Cross-Validation
EEM	Excitation-Emission Matrix
EFSA	European Food Safety Authority
EVD	Eigen Value Decomposition
FRET	Fluorescence Resonance Energy Transfer
HPLC	High-Performance Liquid Chromatography
IARC	International Agency for Research on Cancer
IR	Infrared
JECFA	Joint FAO/WHO Expert Committee on Food Additives
***k*-NN**	*k*-Nearest Neighbors
LC-MS/MS	Liquid Chromatography-Mass Spectrometry / Mass Spectrometry
LV	Latent Variable
MVDA	Multivariate Data Analysis
NIPALS	Non-linear Iterative Partial Least Squares
NIR	Near-Infrared
PC	Principal Component
PCA	Principal Component analysis
PCR	Principal Components Regression
PLS	Partial Least Squares
PLS-DA	Partial Least Squares-Discriminant Analysis
ppb	part per million
ppm	part per billion
RASFF	Rapid Alert System for Food and Feed
RET	Resonance Energy Transfer
RMSE	Root Mean Square Error
RMSECV	Root Mean Square Error in Cross-Validation
RMSEP	Root Mean Square Error in Prediction
SCF	Scientific Committee for Food
SIMCA	Soft Independent Modeling by Class Analogy
SVD	Singular Value Decomposition
UV-Vis	Ultraviolet-Visible

1. INTRODUCTION

According to the European Food Safety Authority, food additives are substances that are intentionally added to foodstuffs to perform desired technological functions, for example, to color, to sweeten and to preserve food. These additives can be either natural or synthetic and its use is regulated by the European Commission (EC No. 1333/2008 [1]).

The food additive *Carthamus* Yellow, extracted from the petals of the Safflower (*Carthamus Tinctorius* L.) is allowed as a coloring foodstuff and as a flavor in the European Union and in the United States according to the EU legislation EC No. 231/2012, however it is not a legal color in food applications in both regions [2]. A coloring foodstuff is a food ingredient with coloring properties that may be added to food and beverages without being declared as E-numbers. It is a compound originated from a natural source and is considered a food ingredient and not additive. [3]

The food sector has been showing an increase in productivity and that has led to the reorganization of the control systems in order to maximize product standardization, guaranteeing a high level of quality and security in food, developing a greater compliance among all batches produced. Moreover, the safety of large scale production sites necessarily passes through systems to highlight possible fraud present throughout the production chain: from the raw materials to the finished products. [4]

In recent years, there have been reported cases of adulteration through the Rapid Alert System for Food and Feed (RASFF) [5] , concerning the contamination of the *Carthamus* Yellow product with the synthetic coloring additive Orange II. On the 2nd of April 2009, a food company from Vietnam and a second company in the United Kingdom reported the case of unauthorized color Orange II in *Carthamus* coming from France, with raw material coming from China. These contaminated products have been redistributed to Vietnam by a food company, which holds the responsibility to analyze the incoming material for posterior distribution to its customers.

The reason for the occurrence of this type of adulteration of the incoming raw materials from the suppliers is the fact that this type of synthetic coloring agents, such as Orange II, has a lower cost. The same adulteration might also happen with natural color ingredients, such as Annatto, which also has a lower cost and a similar color to *Carthamus* yellow.

Currently, in Chr. Hansen, the incoming raw material *Carthamus* Yellow is analyzed by external laboratories, which use Liquid Chromatography-Mass Spectrometry/Mass Spectrometry (LC-MS/MS) technology to detect any undesired substance in the natural coloring product *Carthamus*. The raw material suppliers themselves also claim to analyze the product using High-Performance Liquid Chromatography (HPLC). Both chromatographic methods are time-consuming tasks and, considering that the analysis is done outside Chr.Hansen, it also ends up being an expensive procedure.

In recent years, the laboratories have been employing analytical techniques that are often inadequate because they require many samples, a long time to get the response and staff with high analytical ability. The methods must be easy to use, to promote their use throughout the production chain where it is not always possible to have analytical laboratories. [4]

Therefore, for the purposes of this project, there is a growing interest in developing a fast, cheap and non-invasive method which could identify and possibly also quantify the addition of undesired compounds in the natural *Carthamus* yellow pigment.

Thus, a study was conducted, in collaboration with Chr.Hansen, in which three spectroscopic methods – Ultraviolet-Visible (UV-Vis), Fluorescence and Near-Infrared (NIR) - combined with multivariate data analysis were implemented to develop a method for the identification and quantification of four different unwanted coloring additives – Orange II, Tartrazine, Sunset Yellow and Annatto – present in the *Carthamus* yellow.

2. LITERATURE REVIEW

2.1. Natural coloring product - *Carthamus* Yellow

The compound *Carthamus* yellow, also known as safflower yellow, is the main component of safflower and it is composed of safflomin A and safflomin B, illustrated in Figure 2.1. This substance is obtained by extraction from the petals of safflower (*Carthamus Tinctorius* L.), shown in Figure 2.2, using water or slightly acidized water and, consequently, drying of the extract. [6], [7]

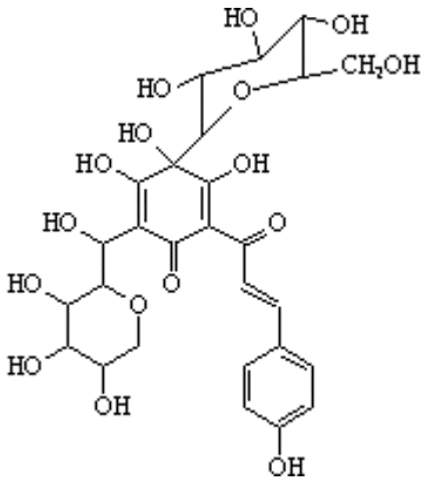
The yellow pigments present in Safflower account for about 30% of its total pigments. Besides safflomin A and B, this plant also contains minor yellow pigments such as precarthamin, safflor yellow A, anhydrosafflor B, safflomin C, tinctomrine and cartormin. A red pigment can also be found but in lower quantity, about 0.83%, however, there is no interest in this compound for the present work. [8], [9]

The main advantage of safflower yellow pigments in food is their high solubility in water, which makes this product rare and valuable in the food industry. Currently, its main application is to make more appealing beverages, dairy products, and confectionaries by being added to juices, yogurt, gelatin desserts, and candy. [8]

As described by *Joint FAO/WHO Expert Committee on Food Additives* (JECFA), *Carthamus* product is a yellow to dark brown compound that can be presented as crystals, paste, liquid or powder and, besides the color pigments, it can also contain sugars, salts and/or proteins, naturally occurring in the source materials. [6] Its physical aspect as powder and liquid is shown in Figure 2.3.

The product is in full compliance with EU Food Regulation 178/2002/EC [10] with later amendments and with EU Regulation 1334/2008 [11] with later amendments on flavorings and certain food ingredients with flavoring properties. *Carthamus* also complies with the specification for identity and purity given by JECFA for *Carthamus*. [12]

Safflomin A



Safflomin B

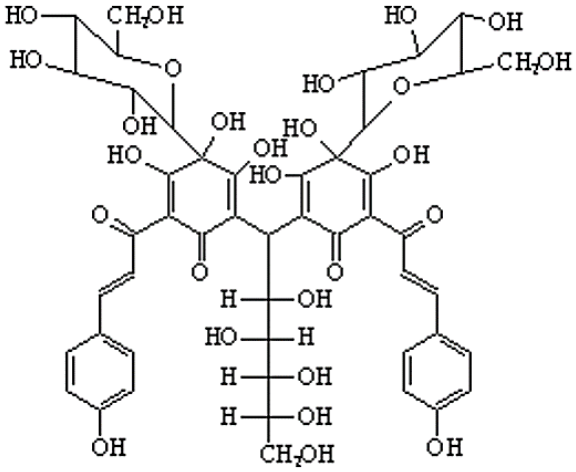


Figure 2.1 Chemical structure of the two main coloring substances present in Carthamus: Safflomin A (Hydroxysafflor yellow A) and Safflomin B (Safflor Yellow B). [6]

Figure 2.2 *Carthamus tinctorius* L. plant, also known as Safflower. [13]

Figure 2.3 Physical appearance of Carthamus in powder (left) and liquid (right) form. [12]

2.2. Unwanted coloring Additives

2.2.1. Synthetic Colors (Azo-Dyes)

Azo dyes, as it name suggests, are the colorant compounds that contain in their chemical structure the azo group ($-N=N-$) which is attached at either side to two sp^2 carbon atoms. The majority of the azo colorants contain a single azo group (monoazo dyes) and, usually, this group is linked to two aromatic ring systems. [14]

Within the category of azo dyes, some of the dyes are classified as illegal for use in foodstuffs in the European Union, according to EFSA review on the toxicology of a number of dyes illegally present in food in the EU [15]. These illegal compounds include the three synthetic dyes addressed in this study - Orange II sodium salt, Tartrazine and Sunset Yellow FCF –, described below.

ORANGE II

The compound Orange II Sodium Salt, also known as C.I. Acid Orange 7, is an orange to orange-brown powder, soluble in water. [16] In Figure 2.4 a) and b), Orange II chemical structure and physical aspect are shown, respectively.

The toxicity of the synthetic compound, Orange II, has been researched by the European Food Safety Authority (EFSA), on the "Review the toxicology of a number of dyes illegally present in food in the EU", that considered that this substance shows some evidence of genotoxicity, although the tests showing these effects are not of a standard protocol and, thus, their relevance to overall risk assessment is not clear. It was also concluded by EFSA that the data regarding its carcinogenicity was not conclusive. [15] The International Agency for Research on Cancer (IARC) did not classify this substance. [17]

According to the RASFF report [5], the maximum amount of this synthetic additive permitted in the natural *Carthamus* product is 500 ppb.

a) b)

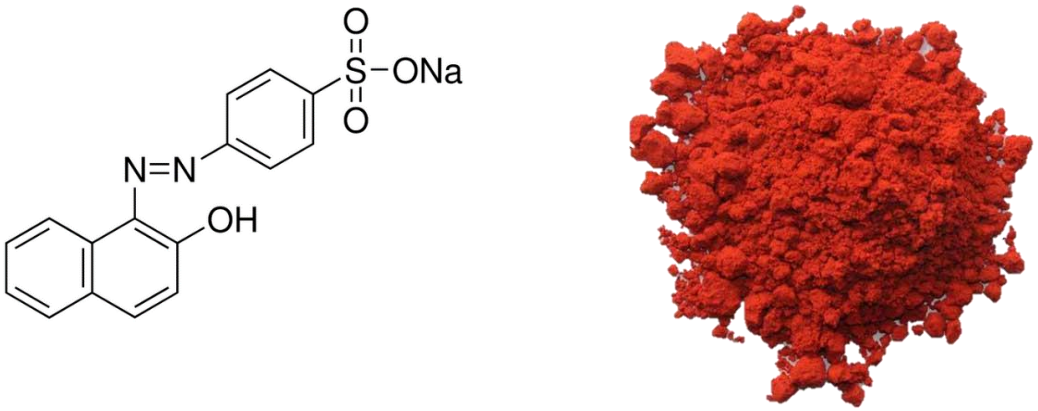


Figure 2.4 Orange II Sodium salt.(CAS: 633-96-5; Molecular Weight: 350.32 g/mol) a) Chemical structure; [16] b) Physical aspect.

TARTRAZINE

The azo dye Tartrazine (*E* 102) is a synthetic compound authorized as a food additive in the EU, soluble in water and sparingly soluble in ethanol. This substance was previously evaluated by JECFA in 1966 and the Scientific Committee for Food (SCF) in 1975 and 1984, in which, both committees have established an Acceptable Daily Intake (ADI) of 7.5 mg/kg body weight (bw)/day. More recently, in 2002, TemaNord recommended that an update of the evaluation of this substance should be made, considering new data from studies on genotoxicity, chronic toxicity/carcinogenicity and reproductive and developmental toxicity. Specifications for Tartrazine have been defined by the EU Commission Directive 2008/128/EC and the Codex Alimentarius. [18]

In Figure 2.5 a) and b), the chemical structure and the physical appearance of this compound, respectively, are displayed.

a) b)

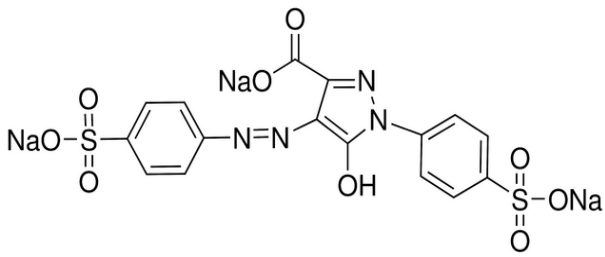


Figure 2.5 Structure of Tartrazine. (CAS: 1934-21-0; Molecular Weigth: 534.36 g/mol) a) Chemical structure; [19] b) Physical appearance.

SUNSET YELLOW FCF

The compound Sunset Yellow FCF (*E* 110), is an azo dye allowed as a food additive in the EU, soluble in water and slightly soluble in ethanol [20]. This compound has been evaluated by JECFA in 1982 and the EU Scientific Committee for Food (SCF) in 1984.The ADI established by these two committees was of 0-2.5 mg/kg bw/day. The most commonly used synonyms in published literature are Sunset Yellow FCF, Food Yellow No. 5, and FD&C Yellow No. 6. .Its chemical structure and its physical aspect are presented in Figure 2.6 a) and b), respectively.

For what concerns the carcinogenicity of this compound, IRCA has classified it as category 3 (“Not classifiable as to its carcinogenicity to humans”). [20]

a) b)

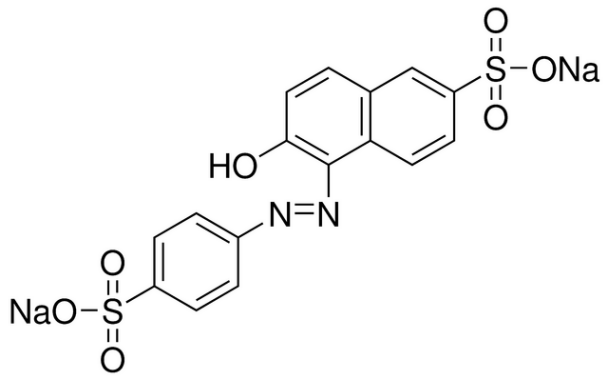


Figure 2.6 Structure of Sunset Yellow. (CAS: 2783-94-0; Molecular Weigth: 452.37 g/mol) a) Chemical structure; [21] b) Physical appearance.

2.2.2. Natural Color Annatto

The natural food color annatto (*E* 160b) is obtained from the outer layer of the seeds of the tropical tree *Bixa orellana* L. presented in Figure 2.7 b). This food additive is permitted by the EU for use in a variety of foods and beverages but not in spices and spice mixtures. Its main coloring constituent is Bixin, with Norbixin being present in smaller amounts. The chemical structure for both pigments is presented in Figure 2.7 a).

The extraction procedure is done by abrading off of the pigment in an appropriate suspending agent for production of the native bixin from the seed or, in alternative, with an aqueous alkaline hydrolysis that, simultaneously, produces norbixin.

For over two centuries, Annatto has been used as food color, primarily in cheese. Currently, its various forms have been used in a wide range of food products. [22], [23]

a) b)

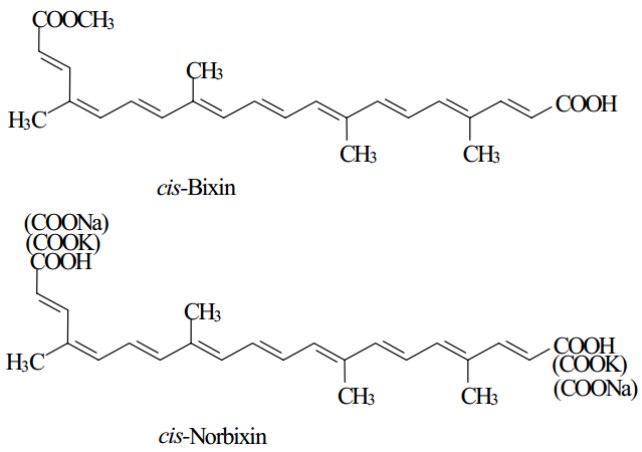


Figure 2.7 a) Chemical structure of the main Annatto pigments, Bixin and Norbixin. Only the cis isomer is shown, although trans isomer is also present; b) Seeds present in the interior of the fruit from the tree *Bixa Orellana* L., from which annatto is extracted. [23]

2.3. Spectroscopy

The development of rapid analytical methods in foodstuffs is based on two approaches: using the substrates physical properties as an information source and the automation of chemical methods. Most of the methods based on physical properties of the food products are spectroscopic methods. [24]

Originally, the term spectroscopy was used to describe a branch of science based on the resolution of visible radiation into its components wavelength. Overtime, the meaning of the term has been expanding including now studies covering the entire electromagnetic spectrum, displayed in Figure 2.8. [25]

Spectroscopy belongs to a group of techniques in which the interaction of matter with electromagnetic radiation gives information about its structure. According to Fessenden & Fessenden, 1993, radiation is characterized by:

- a wavelength (λ) – measured in nm, is the distance between two adjacent maxima;

-a frequency (ν) – measured in hertz, represents the number of oscillations described by the wave per unit of time;

- a wave number (n) - measured in cm^{-1}, represents the number of cycles per centimeter.

As can be seen in Figure 2.8 the entire electromagnetic spectrum is divided into several regions considered as useful for analytical purposes, each region corresponds to the spectroscopic methods associated with these applications and are characterized by a range of wavelengths. [4], [25]

In analytical chemistry the most common spectroscopic techniques are UV-Vis spectroscopy, IR spectroscopy, NIR spectroscopy and Raman spectroscopy. These methods are based upon the phenomena of emission, absorption, fluorescence and scattering. Spectroscopy is extremely useful in analytical chemistry, with applications in a wide number of fields. [25], [26]

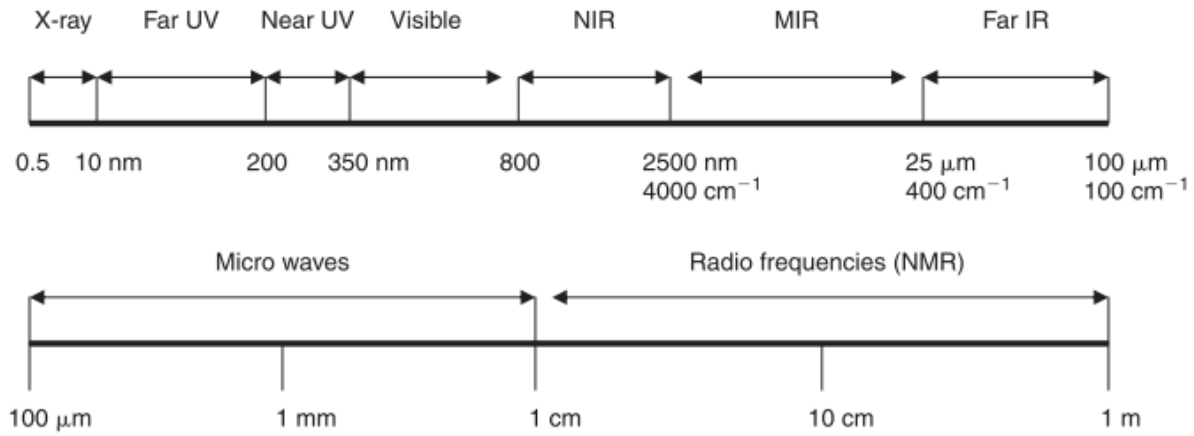


Figure 2.8 Spectral regions of the electromagnetic radiation. [24]

2.3.1. UV-VISIBLE

THEORY AND PRINCIPLES

The UV and visible radiation cover a small part of the electromagnetic spectrum, from 200 nm to 800 nm as seen in Figure 2.8. In UV-Visible spectroscopy, the UV light for the lower wavelengths has the higher energy and, in some cases, this energy is sufficient to cause undesirable photochemical reactions when measuring samples.

Generally, when measuring in UV-Visible spectroscopy, the radiation interacts with matter and multiple phenomena can occur - reflection, scattering, absorbance, fluorescence/phosphorescence (absorption and reemission) and photochemical reaction (absorbance and bond breaking). However, when measuring with UV-Visible spectroscopy, the only desirable phenomenon is absorbance. [27]

Whenever a continuous radiation runs through a transparent material, a fraction of the radiation may be absorbed by the sample. When absorption occurs, if the residual radiation passes through a prism, it displays a spectrum with gaps in it, called absorption spectrum.

Since light is a form of energy, the absorption of light by matter creates an increase of energy content in molecules or atoms and, as a consequence, atoms and molecules change from a state of low energy (ground state) to a state of higher energy (excited state). This mechanism is known as excitation

process and it can be quantified (Figure 2.9). Therefore, the energy absorbed from the electromagnetic radiation is exactly equal to the energy difference between the excited and ground states.

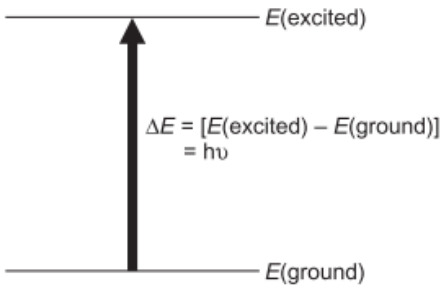


Figure 2.9 The excitation process. [28]

For UV-Vis spectroscopy, the absorption of electromagnetic radiation results in transitions between electronic energy levels. Once a molecule absorbs energy, an electron is promoted from an occupied orbital to an unoccupied orbital of greater potential energy.

For an atom absorbing in the UV-Vis region, the spectrum should show a very narrow absorbance band at wavelengths highly characteristic of the difference in energy levels of the absorbing species (Figure 2.10 a)).

However, for molecules, the absorption usually occurs over a wide range of wavelengths because molecules (as opposed to atoms) normally have many excited modes of vibration and rotation at room temperature. These energy levels are quite closely spaced, corresponding to energy differences considerably smaller than those of electronic levels. Consequently, vibrational and rotational energy levels are superimposed on the electronic energy levels and, since many transitions with different energies can occur, the bands are broadened (Figure 2.10 b)).

Since there are so many possible transitions, each differing just by a slight amount from the remaining ones, each electronic transition consists of a vast number of lines so close to each other that the spectrophotometer is not able to resolve them. In these types of combined transitions, the UV spectrum of a molecule usually consists of a broad band of absorption centered near the wavelength corresponding to the major transition. [27], [28]

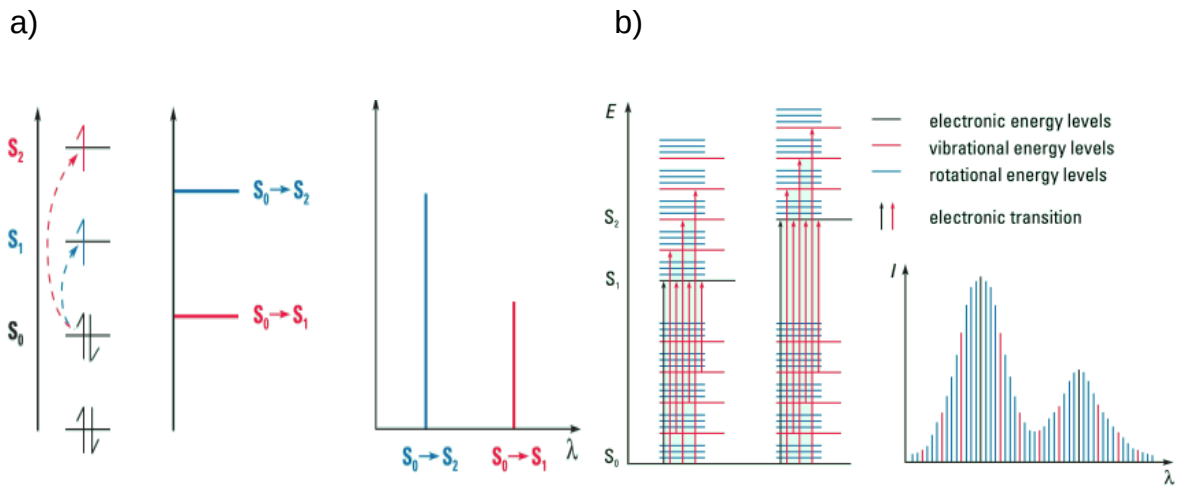


Figure 2.10 a) Electronic transitions and spectra of atoms; b) Electronic transitions and UV-visible spectra in molecules. *[27]*

The greater the number of molecules able to absorb light for a given wavelength and the more effectively these molecules are absorbing, the greater the extent of light absorption. Having this in consideration, the Beer-Lambert law was formulated. This empirical expression (shown below) shows that the absorbance of a solution is directly proportional to the concentration of the sample.

$$A = log\left(\frac{I_0}{I}\right) = \varepsilon\, c\, l \qquad \text{Equation 2.1}$$

Where the term *log (I_0/I)*, also represented by *A*, is the absorbance of the solution (or the optical density in older literature); *c* is the concentration and *l* is the path length of the sample. The molar absorptivity, given by the Greek symbol epsilon, ε, is not a function of the parameters involved in the sampling preparation since it is a property of the molecule itself when undergoing an electronic transition. In practice, when using the Beer-Lambert law to analyze a solution of unknown concentration, solutions of known concentration need to be prepared, a suitable band chosen and the absorbance at this wavenumber measured, in order to create a calibration graph. Therefore, if only one compound is absorbing it is possible to read the concentration of the compound in solution, given that its absorbance is known. If more compounds are absorbing, several wavelengths must be measured and, subsequently, multivariate data analysis should be used in order to resolve the spectra.

However, this law may not be obeyed when there is an equilibrium between the different forms of the absorbing molecule, when the solute together with the solvent forms complexes, when there is a thermal equilibrium between the ground state and the low-lying excited state or when fluorescent compounds changed by irradiation are present. [28], [29]

Instrumental Design

A spectrophotometer is an instrument used to measure the absorbance or transmittance of a sample according to the wavelength of the electromagnetic radiation. [27] The usual UV-Visible spectrophotometers are composed of a light source, a monochromator and a detector. [28] The main elements of the UV-Visible spectrophotometer are described below, according to [27], [28], [30]:

- **Light source:** The first light source normally used is a deuterium lamp, able to emit a good intensity continuum in the ultraviolet region and a useful intensity in the visible region (180 to 350 nm). The second light source is a tungsten-halogen lamp, used to emit a good intensity radiation over part of UV and over the entire visible region of the spectrum (330 to 900 nm). The majority of the spectrophotometers used to measure in the UV-Visible region contain both lamps and it can either have a source selector able to switch between the lamps as appropriate, or the light from the two sources is mixed to produce a single broadband source.
- **Monochromator:** is the component of the spectrophotometer responsible of spreading the beam of the light into its component wavelengths. In common spectrophotometers the monochromators have a prism or a diffraction grating. Through a system of slits it is possible to focus in the desired wavelength on the sample cell. Consequently, the light passes through the sample cell and reaches the detector, where the signal is recorded. When a higher

resolution is required, a double monochromator can be used. This component improves the stray light rejection and allows measurement of sample with high optical density. However, double monochromators have a lower optical throughput, which causes the signal to noise ration to deteriorate.

- **Detector:** The detector can be either a photomultiplier tube or a photodiode detector and its function is to convert a light signal into an electrical signal. Photodiodes are smaller and cheaper, whereas photomultipliers have a higher sensitivity. In typical double-beam instruments, the light emitted by the light source is split into two beams, the sample beam and the reference beam. When the sample cell is empty in the reference beam, it is assumed that the detected light is equal to the intensity of light entering the sample.

SAMPLING

In an experiment, the sample is measured in a cell that must be made of a material transparent to the electromagnetic radiation in use. Even though cells composed of glass or plastic are appropriated for measurements in the visible region, when measuring in the UV region a cell made of quartz must be used, since glass and plastic absorb this type of radiation.[28] The cells used in UV-Visible spectroscopy are commonly used to measure liquids or solutions. For quantitative analysis purposes, transmittance measurements using samples in liquid form or solutions is simpler and more accurate than measuring reflectance in solid samples. [27]

In order to measure transmittance, the sample is prepared mostly by dilution with a suitable solvent. The most commonly used solvents are water and ethanol since they are both cheap and transparent down to about 210 nm. However, for less polar samples, hexane and other hydrocarbons are more suitable.[30]

2.3.2. FLUORESCENCE

THEORY AND PRINCIPLES

The use of fluorescence in biological sciences have been growing significantly in the past years and Fluorescence spectroscopy is now used in biotechnology, flow cytometry, medical diagnostics, DNA sequencing, forensics, and genetic analysis.

Fluorescence spectroscopy has the advantage of being a more selective and sensitive method when compared to absorption spectroscopy, as its spectral information is simpler, with fewer fluorophores and, thus, the detection of trace compounds is improved. Also there is no longer the need for the expense and difficulties of handling radioactive tracers for most biochemical measurements. Thus, this spectroscopy is now very used for many scientists in different disciplines. [31], [32]

Luminescence is the emission of ultraviolet-visible or infrared photons from electronically excited species. The two most common types of luminescence are fluorescence and phosphorescence. The classification of the type of luminiscence depends on the nature of the excited state as can be see in Table 2.1. [32], [33]

Table 2.1 Different types of luminescence. [33]

Phenomenon	*Mode of excitation*
Photoluminescence (fluorescence, phosphorescence, delayed fluorescence)	Absorption of light (photons)
Radioluminescence	Ionizing radiation (X-rays, α, β, γ)
Cathodoluminescence	Cathode rays (electron beams)
Electroluminescence	Electric field
Thermoluminescence	Heating after prior storage of energy (e.g. radioactive irradiation)
Chemiluminescence	Chemical process (e.g. oxidation)
Bioluminescence	Biochemical process
Triboluminescence	Frictional and electrostatic forces
Sonoluminescence	Ultrasounds

A fluorophore is a chemical compound similar to a chromophore that emits photons. Normally, fluorescent compounds have several combined aromatic groups and double bounds in its chemical structure. In Figure 2.11, some common fluorescent molecules are shown. The fluorescence properties of these compounds depend on their structure and on their environment. Fluorophores can be divided into two groups: intrinsic and extrinsic fluorophores. The first one relates to the compounds that occur naturally in molecules, such as aromatic amino acids, NADH, flavins, derivatives of pyridoxyl, and chlorophyll. On the other hand, extrinsic fluorophores are added to the sample in order to provide fluorescence when nonexistent or to change some spectral property of the sample. These include dansyl, fluorescein, rhodamine, and numerous other substances. Even though these compounds are the most commonly known fluorescent molecules, there is a vast range of fluorophores. [32], [34]

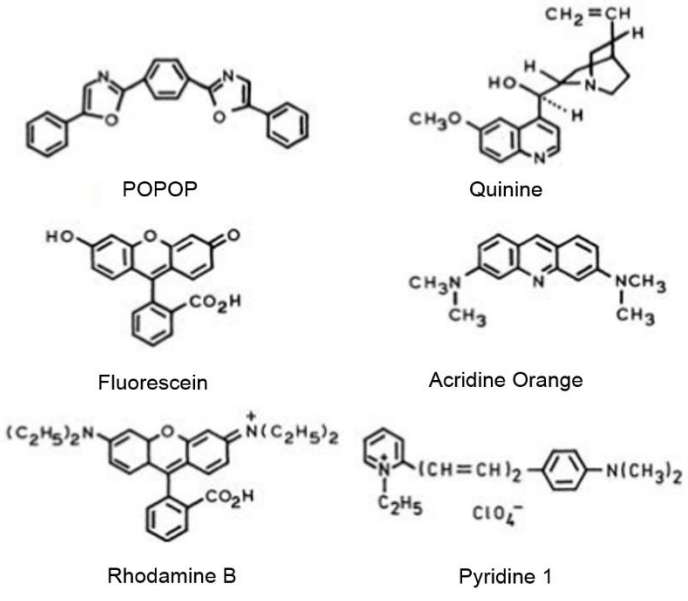


Figure 2.11 Structures of common fluorescent substances. [32]

The Fluorescence spectroscopy handles the excitation and emission in molecules. Every molecule is capable of going into an electronically excited state when exposed to light of a wavelentgh (energy level) equal to the energy gap between the ground state and the excited state, this process is known as molecular absorbance of light. As seen before for UV-Visible spectroscopy, the amount of

light absorbed is proportional to the concentration of the absorbing molecule, according to the Beer-Lambert law (Equation 2.1). [35]

The phenomenon of absorbance only consists in transitions from the ground state S_0 to the excited state S_n (n>1). Then an excited molecule will return to the ground state S_0 through the following sucessive steps:

- The molecule present at an excited state S_n returns to the lowest excited state S_1 through the dissipation of a part of its energy to the nearby environment. This process is commonly known as internal conversion.
- From the excited state S_1, some molecules are capable of returning to the ground state S_0 by different phenomena. One of them, is through the emission of a photon with a radiative rate constant k_r. This phenomenon is called fluorescence.

In Figure 2.12, the Jablonski diagram, also known as electronic transitions diagram, is displayed, which illustrates the mechanism of the excitation/relaxation occurig in the molecule. [34]

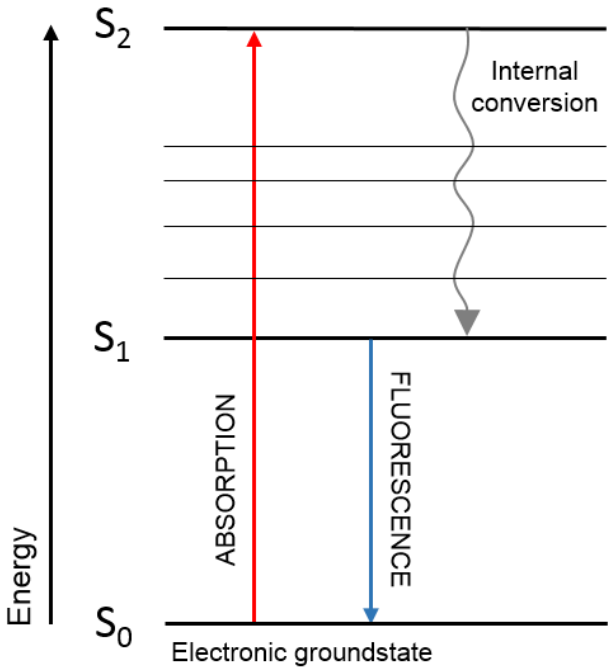


Figure 2.12 Example of a Jablonski diagram. Adapted from [30].

In principle, the relaxation through emission of light always occur from the lowest energy excited electronic state of a molecule (S_1), however, when a molecule is excited to a higher energy state (S_2 or S_3), the molecule goes through internal conversion, which precedes emission. By analyzing the Jablonski diagram, it is possible to verify that the energy of the emission is normally less than the energy of the absorption and, therefore, the emission spectrum has its maximum shifted to longer wavelengths compared to maximum absorption spectrum.

When an absorption and/or emission spectra of a fluorophore has two or more bands, there is a shift corresponding to the difference between the two most intense bands of these two spectra. This difference is known as Stokes shift and is expressed in wavenumbers. This parameter is able to provide

information on the excited states and, thus, if this shift is larger it is easier to detect fluorescent species. [32], [34]

According to Kasha's rule, the same fluorescence emission spectrum generally is observed regardless of the excitation wavelength. When there is excitation into higher electronic and vibrational levels, the excess of energy is rapidly dissipated, leaving the fluorophore in the lowest vibrational level of S_1. Due to this rapid relaxation, the emission spectra are normally independent of the excitation wavelength. There are some exceptions, for instance, when a fluorophore exist in two ionization states, they show different absorption and emission spectra.

Typically, in the fluorescence process, the return to the ground state occurs to a higher excited vibrational ground state level and, as a consequence, the emission spectrum of the molecule is normally a mirror image of the absorption spectrum of the S_0 to S_1 transition. This symmetry happens since the electronic excitation does not significantly alter the nuclear geometry and, thus, the spacing of the vibrational energy levels of the excited states are similar to that of the ground state. Therefore, the vibrational structures seen in the absorption and the emission spectra are similar. This property in the fluorophore is known as the mirror-image rule, although it must be noted that not all compounds comply with this rule.

Usually, the data obtained from Fluorescence spectroscopy is presented as emission spectra. This spectrum is measured as the light emitted (fluorescence) across a wide range of wavelengths upon excitation at a fixed wavelength, giving a plot that displays the fluorescence intensity versus the emission range, in terms of wavelength (nm) or wavenumber (cm^{-1}). Alternatively, an excitation spectrum can also be obtained by measuring the emission at one fixed wavelength while exciting the molecule over a wide range of wavelengths. [32], [35]

However, in Food Science analysis it is very common to use fluorescence Excitation-Emission Matrix (EEM), also known as fluorescence landscape, which is further investigated through multivariate data analysis. In order to obtain the fluorescence data in landscape, several emission spectra at different excitation wavelengths (or vice versa) must be measured, thus creating an excitation-emission map that covers the total area of fluorescence. This structure has the advantage of detecting analytes or interferences present in different areas. [31], [35]

Resonance Energy Transfer

For the purpose of the present study, the process of resonance energy transfer (RET), also known as Fluorescence or Förster resonance energy transfer (FRET), is studied. This non-radiative phenomenon relies on the interaction between a donor molecule in the excited state and an acceptor molecule in the ground state. This phenomenon occurs when the emission spectrum of a donor (not necessarily a fluorescent compound) overlaps the absorption spectrum of another substance (acceptor), as illustrated in Figure 2.13. Due to this overlapping, several vibrational transitions in the donor have basically the same energy as the corresponding transitions in the acceptor. The energy involved in these transitions is transferred by resonance, in which the electron of the excited molecule induces an oscillating electric field that excites the acceptor electrons and, therefore, the acceptor reaches an excited state.

There is no intermediate photon in RET and, thus, the donor and the acceptor are coupled through dipole-dipole interactions. The rate of energy being transfer is subjected to the degree of overlap of the donor emission spectrum with the acceptor absorption spectrum, the quantum yield of the donor, the relative orientation of the donor and acceptor transition dipoles, and the distance between the two interveners. [32]–[34]

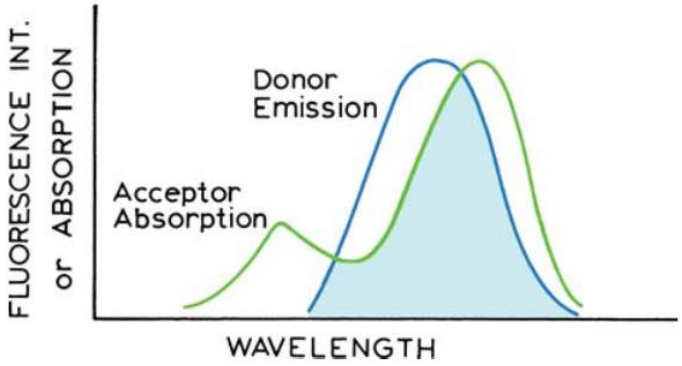


Figure 2.13 Spectral overlap related to the resonance energy transfer (RET). [32]

INSTRUMENTAL DESIGN

An instrument for measuring fluorescence is called a spectrofluorometer or a fluorescence spectrophotometer. In Figure 2.14, a schematic diagram of a conventional spectrofluorometer is displayed. The main components of this instrument are presented and described below:

- **Light source:** Normally, a xenon lamp is used as a source of emitting light, since it has a high intensity at all wavelengths ranging upward 250 nm. Other light sources are also used, such as Pulsed Xenon Lamps, High-Pressure Mercury (Hg) Lamps, Xe–Hg Arc Lamps, among others, however, this components are only used in very special spectrofluorometers and, thus, they are not used very often.
- **Monochromator:** A monochromator is used to disperse polychromatic or white light into various colors or wavelengths. In spectrofluorometers this component is used to select both excitation and emission wavelengths. Therefore, there is an excitation monochromator and an emission monochromator. The features that influence the performance of a monochromator are the dispersion, efficiency and the stray light levels. Stray light means the light is transmitted in wavelengths outside the chosen wavelengths.

  One of the factors influencing the choice of a monochromator is the efficiency to maximize the ability to detect low light levels. On the other hand, the resolution is not very important since the emission spectra hardly have peaks with line widths less than 5 nm. A monochromator normally has an entrance and exit slit. Larger slits widths increase signal levels and, consequently, higher signal-to-noise ratios. Smaller slit widths increases the signal level, however the light intensity decreases.

The dispersion can be achieved using prisms or diffraction gratings, the latter being the most commonly used. Imperfections in the gratings are the cause by stray light transmission. Monochromators can either have plane or concave gratings. In the spectrofluorometer schematized in Figure 2.14, the excitation monochromator has two concave gratings. The concave gratings are produced by holographic and photoresist methods and they can have fewer reflecting surfaces, lower stray light and can be more efficient. Thus, concave gratings can serve as both the diffraction and focusing element, resulting on one element instead of three reflecting surfaces. Therefore, these are more used comparing to the plane version.

In general, the excitation monochromator is chosen for high efficiency in the ultraviolet wavelengths and the emission monochromator for high efficiency at the visible wavelength range. The automatic scanning of wavelengths in spectrofluorometers is possible by the motorized monochromators, which are controlled by the electronic devices and the computer (where the data is stored).

- **Optical filters:** These components can be used to compensate the flaws in the behavior of the monochromators. In addition, when the spectral properties of a fluorophore are known, filters are a more viable option to obtain maximum sensitivity rather than using monochromators.
- **Detector:** Almost every spectrofluorometer uses a photomultiplier tube (PMT) as the detector of the instrument. A PMT works as a current source, and this current is proportional to the light intensity. This component responds to individual photons and the pulses can be detected as an average signal or counted as individuals.
- **Optical module:** As can be seen in the instrument schematic, the optical module surrounds the sample holder and its main components are a shutter, a beam splitter, a reference cell and a polarizer. Shutters are used to eliminate the exciting light or to close off the emission channel. Beam splitters is a thin piece of clear quartz used in the excitation light path used to reflect part of the excitation light to a reference cell, in which there is a stable reference fluorophore. Polarizers are removable pieces of the instrument and they are normally inserted only for measurements of fluorescence anisotropy or when it is necessary to select particular polarized components of the emission and/or excitation light path. In the schematic diagram of the instrument there is also an additional light path in the right, used to measure fluorescence anisotropy by the T-format method. However, nowadays, with modern electronics, there is no need to use this method.
- **Sample holder:** This module of the instrument must be versatile since the position of the sample can change depending on the purpose of the measurements. [32], [34]

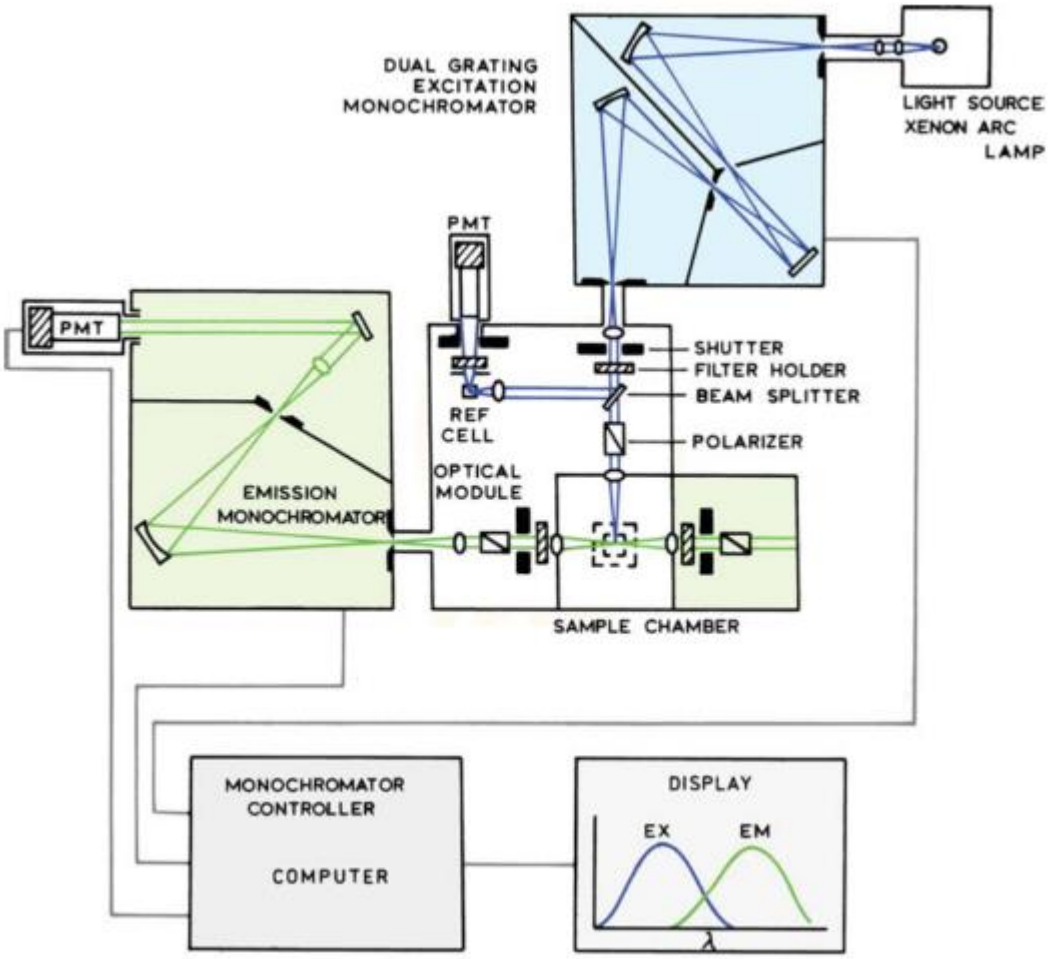


Figure 2.14 Schematic layout of a spectrofluorometer. [32]

MEASURING METHODS AND SAMPLING

The geometry of the sample illumination and the optical density of the sample have a great impact on the apparent fluorescence intensity and spectral distribution of a sample. Commonly, the geometry used in Fluorescence spectroscopy is the right-angle observation of the center of a centrally illuminated cuvette (Figure 2.15 a). There are also other geometric arrangements used, including front-face and off-center illumination. However, for the purposes of the present thesis, only the right-angle and front-face method (Figure 2.15 b)) will be explored. [32]

In cases where the sample is very thick and dense, the measurements using a right-angle are difficult to perform. Actually, high optical densities at excitation and emission wavelengths may decrease the real fluorescence intensity of the sample and also distort the fluorescence spectra. With the usual right-angle observation, the instrument only detects the fluorescence emitted from the central part of the exciting beam and, therefore, when the concentration of the sample is high, a significant part of the incident light is absorbed before reaching the central part of the cuvette. This is known as the excitation inner filter effect.

To overcome this problem, front-face Fluorescence spectroscopy can be used. By using this method, the sample is excited at the cuvette surface, in a way that any displacement of the excitation light through the sample to the cuvette center is avoided. Therefore, the fluorescence spectra (in excitation and emission) using front-face mode is not distorted. Using this technique the excitation light is focused to the front surface of the sample, and the fluorescence emission is collected from the same region, using an angle that decreases the reflected and scattered light. This technique is adequate for highly concentrated and opaque or solid samples. [33], [34]

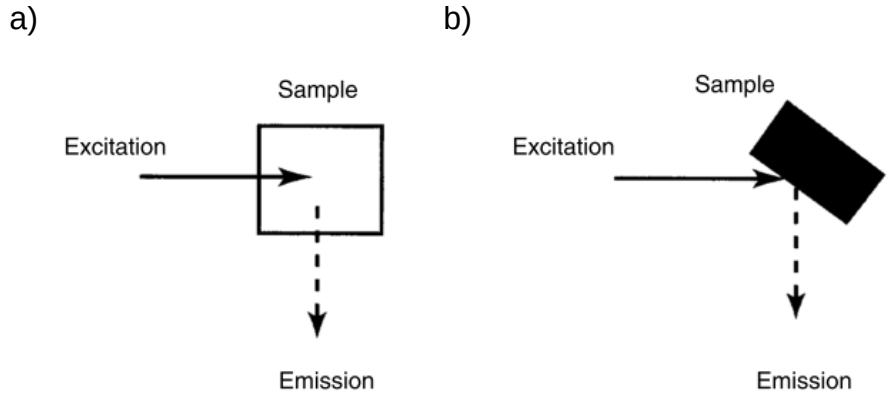


Figure 2.15 Schematic representation of two geometric arrangements for observation of Fluorescence spectroscopy: a) right-angle and b) front-face. [34]

Fluorescence Data Pre-Processing

Before using the fluorescence data in multivariate data analysis there is the need to remove all the areas that do not correspond to useful fluorescence information. These regions on the fluorescence landscape are normally Rayleigh scattering peaks, first and second order, and variables lower than the first order Rayleigh peak, but also areas with no information at all. In addition, the area above the second order Rayleigh peak only holds the same information which already is present between the Rayleigh peaks. These regions are removed in order to create a simpler model.

There is also the possibility of having a Raman peak that is usually hidden below or is slightly visible in the spectra. Subtracting a blank spectrum is one way to remove the Raman scatter.

After the removal of all the unwanted areas, the spectral area is reduced, including only the useful data to construct the chemometric model. Figure 2.16 shows an example of a raw fluorescence landscape obtained by measuring a diluted sample in front-face mode and the landscape after removing the unwanted areas (Figure 2.17).

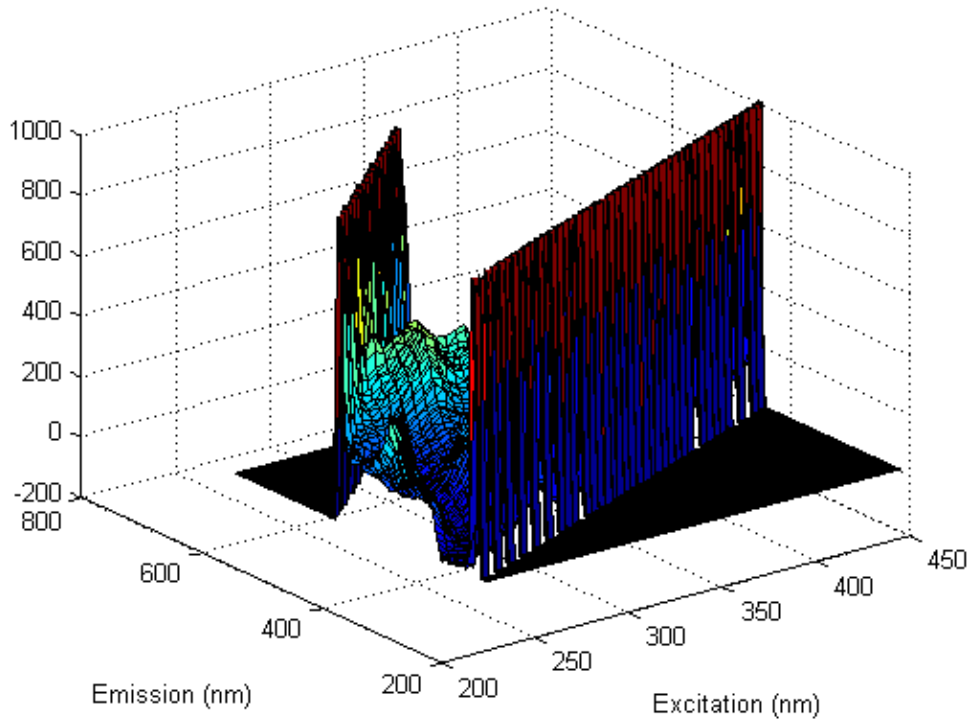


Figure 2.16 Example of a fluorescence landscape before pre-treatment (Visualization in Matlab).

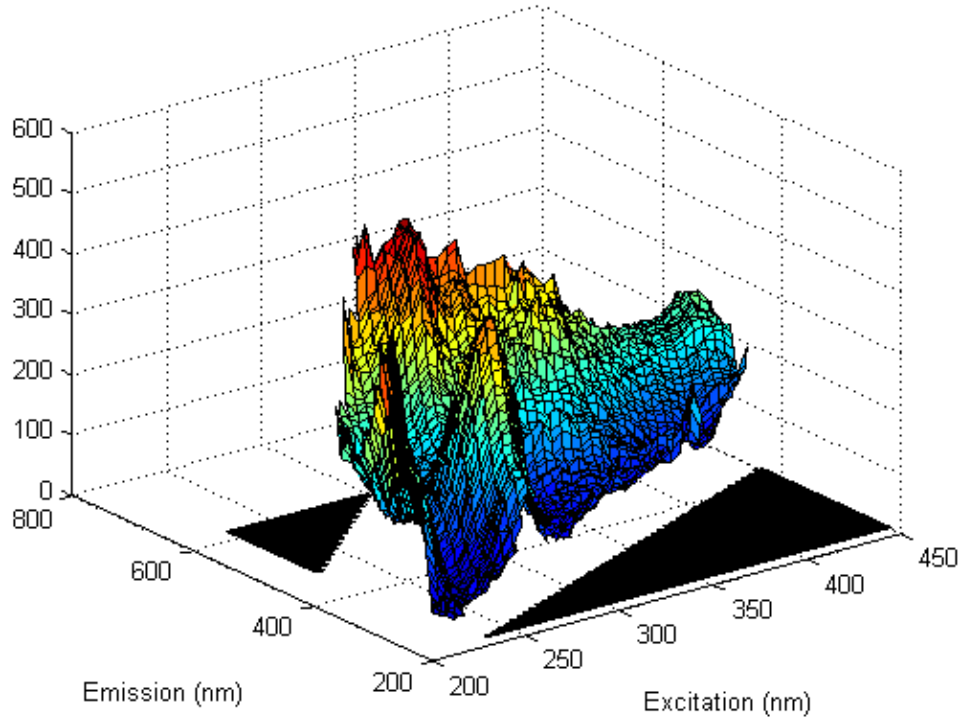


Figure 2.17 Example of a fluorescence landscape after the pre-treatment (Visualization in Matlab).

2.3.3. NEAR-INFRARED

In the last ten years, NIR spectroscopy coupled with chemometric techniques has become an emerging analysis method for control and monitoring in various industries. The interest in NIR spectroscopy has been increasing thanks to the improvement of the instruments together with the computer progresses and development of new mathematical models and modelling techniques to deal with the data obtained. [36]–[38]

NIR comprises the region between the visible and the mid infrared (MIR) regions of the electromagnetic spectrum. The spectra in the NIR region (800–2500 nm, respectively 12821–4000cm^{-1}) normally results from absorption bands related with the vibration and combination overtones. These bands are rather weak due to the intensity loss for each step from the fundamental to the next overtone. NIR spectroscopy requires a dipole moment change and anharmonicity of the vibrating atoms of the molecules. In NIR spectra, X-H stretching or combinations of stretching and bending vibration overtones are very typical. The first overtones of most of the X-H fundamentals absorb at wavenumbers less than 2000 cm^{-1} and hence they appear in the NIR frequency range. NIR present as characteristics a low molar absorptivity and scattering. [24], [28], [30], [38]

This technique is widely used, due to its speed, low cost and easy (and non-destructive) sampling. However, the spectral bands of polyatomic molecules display many overtones and combination vibrations that overlap and make a typical NIR band look very broad and featureless, making it difficult to interpret. In order to solve this, the NIR spectra should be assigned with reference to their molecular origin and, subsequently, combined with chemometric evaluation techniques, giving a more effective application for research purposes. This chemometric analysis allows a better understanding of the NIR data. [29], [30], [36], [38]

The instruments used in NIR spectroscopy are not very distinct from visible or infrared spectroscopy. In some cases, conventional instrumental design also used for the visible or IR region can be used to measure in NIR. However, the main distinction of NIR spectrophotometers to the remaining ones is their specialized applications. NIR spectroscopy does not require such a high resolving power, like visible, UV or infrared spectroscopy, since NIR spectra shows multiples of combination tones and overtones of absorption lines. The preparation of samples in NIR spectroscopy is not very demanding, comparing to other spectroscopic methods. Nowadays, the sample holders are very versatile and adequate to any type of sample. [38]

2.4. Chemometrics / Multivariate Data Analysis

Chemometrics can be defined “as the chemical subject that uses mathematics, statistics and formal logic to (a) design or select optimal experimental procedures; (b) provide maximum relevant chemical information by analyzing chemical data; and (c) obtain knowledge about chemical systems.”. [39]

Even though the definition of chemometrics is quite broad, its most relevant tool is the application of multivariate data analysis (MVDA) in data acquired from analytical chemistry methods. The MVDA has been established as a powerful technique to analyze and structure data sets related to chemistry and biochemistry areas. [40]

In scientific disciplines like chemistry, essentially analytical chemistry, and food science, the data acquired from the instrumental measurements is quite complex, often consisting thousands of variables for each sample. The complexity of the data has led to the need of multivariate data analysis to resolve the information in the data. [41]

The experimental data can be evaluated through quantitative and qualitative analysis. The methods used for qualitative analysis are divided into two groups: supervised and unsupervised learning. The latter approach does not require additional information about the samples to be classified, and it is employed for exploratory data analysis or for empirical investigation of samples. On the other hand, the supervised learning approach assigns new objects to already established classes. [30], [38]

Figure 2.18 displays a diagram with the most commonly used methods for data analysis according to the purpose of the analysis.

For the purpose of this project, Principal Component analysis (PCA), Partial Least Squares (PLS) regression, Partial Least Squares-Discriminant Analysis (PLS-DA), Soft Independent Modeling by Class Analogy (SIMCA) and *k*-Nearest Neighbors (*k*-NN) are presented and described in the following subsections.

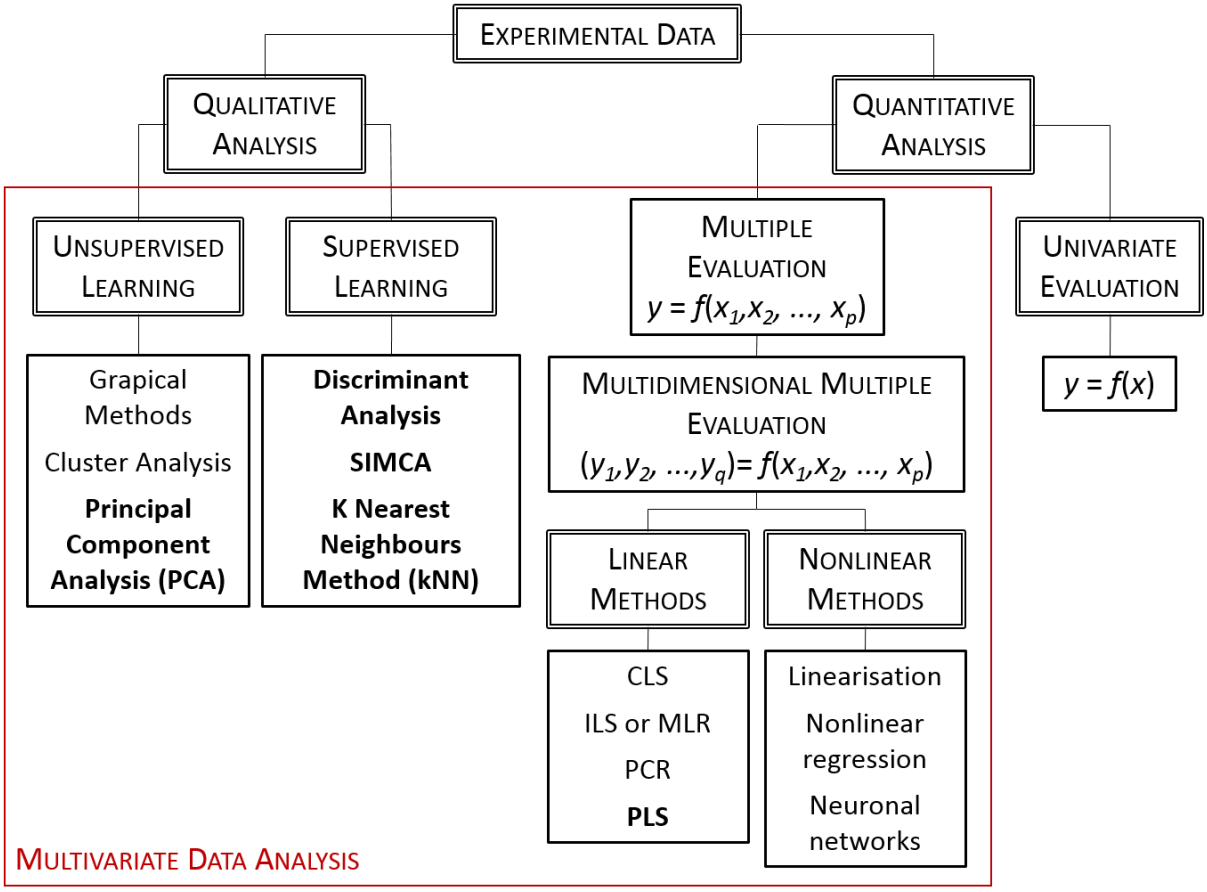


Figure 2.18 Diagram showing the various methods used in data analysis. The methods that appear in bold are the ones used in this study. Adapted from [30].

2.4.1. Principal Component Analysis (PCA)

Theory and Principles

PCA is a bilinear projection and decomposition technique which is able to reduce a data set matrix **X** (*K x N*) (Figure 2.19), of large dimensions to a much smaller number of *A* variables, called principal components (PCs). These components capture the similarities/dissimilarities between the samples and variables constituting the modelled data. PCA has the valuable feature of allowing a simultaneous and interrelated view over both, samples and variable spaces, by performing a linear transformation under the constraints of keeping data variance and imposing orthogonally of the PCs.

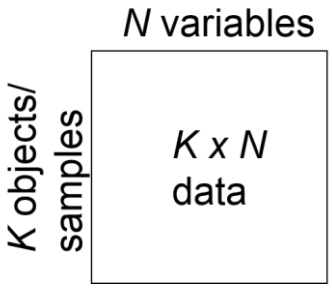


Figure 2.19 The data matrix ***X***.

PCA constructs its mathematical model by decomposing the data matrix **X** using the following expression:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \text{Equation 2.2}$$

Where **T** (*K x A*), **P** *(N x A)* and **E** *(K x N)* are the PCA scores, loadings and residuals matrices, respectively. The scores represent the coordinates of samples in the PC space, thus score plots allow the inspection of sample similarities and differences and makes the detection of sample groupings and trends easier. The loadings give the weight with which each original variable contributes to the PCs, therefore, through a loading plot it is possible to inspect the correlation structure among the variables, hence variables that present equal or close values in the loadings, are correlated (anti-correlated, if they are on opposite sides of the origin). The residuals matrix, also known as the noise or error matrix and with the same dimensions as **X**, contains the part of the data which is not explained by the model. This matrix is often used to identify outlying samples and/or variables. [41], [42]

The principal components consist of linear combinations of the initial variables. They are orthogonal to each another and are computed in such a way to cover the largest amount of variance. The first PC is the linear combination that captures as much as possible of the variability in all the original variables. The second PC corresponds to the normalized linear combination that has most of the remaining variables and is orthogonal (uncorrelated) to the first PC. Each successive PC accounts for as much of the remaining variability as possible until all principal components have been calculated. [43], [44] The orthogonality of PCs vectors causes the complete elimination of correlation present in the data set by using the new variables (PCs) instead of using the original **X**. Therefore, from a geometrical

point of view, PCA is an orthogonal projection/linear mapping of the data set **X** in the coordinate system generated by the loadings **P**.

The purpose of PCA is to identify directions of the data that permit a reduction of the data to a simpler and smaller data matrix by eliminating useless information. In order to achieve this, PCA uses different mathematical algorithms to calculate the eigenvectors and eigenvalues of the data matrix. The main algorithms used - EigenValue Decomposition (EVD), POWER method, Singular Value Decomposition (SVD) and Non-linear Iterative Partial Least Squares (NIPALS) - vary depending on the matrix used to work on and whether the PCs are obtained simultaneously or sequentially. The most commonly used algorithms are NIPALS and SVD. [41], [42]

Selection of the number of Principal Components

A PCA model is only determined once the number of PCs is fixed. The user is the one that chooses the number of PCs used, which is always much smaller than the number of original variables and objects. The choice of the optimal number is subjective and it should depend on the need to explain the variance in the original data but, at the same time, avoiding over fitting. Therefore, the data is not compressed until the number of PCs is decided.

Even though there are an extensive number of methods to choose the optimal number of PCs based on formal test of hypothesis, in MVDA it is important not to assume that PCs follow a specific distribution. Therefore, a more practical approach is preferable, especially graphical visualization criteria, such as sequential exploration of scores plots and/or inspection of the residual plots, plots of eigenvalues (scree plots) or cumulative variance versus the number of components.

It is possible to evaluate the optimal number of PCs through the examination of the scores plots for each component sequentially, based on the data pattern present in the plot and stopping when there is no more relevant information in the pattern, making the present number of components optimal.

It is also possible to examine the plot for the cumulative variance versus the number of components to verify the optimal number of components. This evaluation consists on choosing the number of PCs that corresponds to a change in the slope of the curve present in this plot from steep to shallow.

Another simple technique is to choose the number of components that gives a percentage of accounted variance of 80-90%. In addition, by analyzing the residuals plots it can be observed if there is some systematic variation that is unmodelled. [30], [41], [42]

Outlier Detection

PCA is quite sensitive to outliers - values which are not representative for the rest of the data - since the directions of the PCA performed are influenced by outliers. As matter of fact, outliers artificially increase the variance in an uninformative direction, driving a PC along, which is not favorable for the viability of the model created. Even though the method used to detect outliers depends on the nature of the data, outliers are mostly detected by analyzing strangely distant objects in the score plot, in the residuals plot or in the Hotelling's T^2 plot - which gives the distance of a sample from the center of the PC's plane. [41], [44], [45]

Q, also known as *DModX*, gives the measure of the distance of a sample from the PCA model, corresponding to the sum of squares of each sample (row) of **E,** as follows:

$$\boldsymbol{Q}_i = \boldsymbol{e_i} \cdot \boldsymbol{e_i^T} \qquad \text{Equation 2.3}$$

Where $\mathbf{e}_i$ is the *i* th row of **E**. The *Q* values obtained are a measure of the difference, or residual, between a sample and its projection onto the *A* PCs retained in the PCA model, revealing how well each sample complies with that model.

On the other hand, the distance of a sample from the center of the PC's plane is known as T^2 or Hotelling's T^2 statistics, and it refers to the sum of the normalized squared scores. T^2 is the measure of the variation in each sample within the PCA model and is given by:

$$\mathbf{T}_i^2 = \boldsymbol{t_i}\, \boldsymbol{\lambda^{-1}}\, \boldsymbol{t_i^T} \qquad \text{Equation 2.4}$$

Where $\mathbf{t}_i$ refers to the *i* th row of the scores matrix, **T,** and **λ** is a diagonal containing the eigenvalues (λ_1 through λ_A) corresponding to the *A* eigenvectors (PCs) retained in the model.

By plotting the two distances described above, *Q* and Hotelling's T^2, the T^2 versus *Q*-residual plot is obtained, from which is possible to inspect peculiar samples, such as outliers. [41]

2.4.2. Partial Least Squares (PLS) Regression

Theory and Principles

The PLS regression is a technique used to relate two sets of variables, *X* and *Y*, through an auxiliary set of variables known as latent variables (LVs), or PLS factors or components, which are linear combinations of the variables x_1, x_2, x_3,…, x_k, and very similar to the components used in PCA and in other methods, such as Principal Components Regression (PCR). [42]

The purpose of PLS is to find a small number of factors *A* that are predictive for the **Y**-block data and make use of the **X**-block data efficiently, through a decomposition of **X** into a set of orthogonal factors, which are used for fitting **Y**. [45] PLS ensures that the first latent variables contain the maximum predictive information as possible and it uses the dependent variable *Y* in the data compression and decomposition operations. Hence, once the first component is computed, a deflation step is necessary, to eliminate from both **X** and **Y** the portion of variation already accounted for.

This technique uses both **X** and **Y** data actively in the data analysis in such a way that most variance in both **X** and **Y** is explained. This feature of PLS allows the minimization of the potential effects of X-variables having large variances not relevant for the calibration model. Indeed, PLS aims to find components that compromise between explaining the variation in the **X**-block and predicting the responses in **Y**, and each component is obtained by maximizing the covariance between **Y** and all possible linear functions of **X.** [42], [44], [46]

The algorithm used in PLS consists of a mix of the two PCA computations, one for **X**-block and one for **Y**-block, using the NIPALS algorithm. Other algorithms are also commonly used, such as

SIMPLS. [42], [45] According to [47], concerning a study comparing 9 different PLS algorithms (including NIPALS and SIMPLS), the NIPALS algorithm, developed by Wold [48], is numerically stable for both low and high number of PLS latent variables, however, it is one of the slower algorithms, which is caused by the deflation of **X** and the calculation of two sets of loading vectors required in this algorithm. The SIMPLS algorithm, developed by de Jong [49], can also be numerically stable but for a reasonable number of PLS factors. If a higher number of LVs is used it shows a tendency towards numerical instability. The use of high number of factors might not be relevant and, thus, SIMPLS might be considered as relevant, however, the tendency of becoming more unstable, which relates to the degree of orthogonally of its score vectors, was evident in this study. [50]

The same way as in PCA, the matrices **X** and **Y** are centered or auto scaled before the decomposition into factors. Considering *K* samples, *A* factors, *N* variables and *P* analytes, a PLS bilinear model can be represented mathematically as follows:

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P^T} + \boldsymbol{E} = \sum_{a=1}^{A} \boldsymbol{t_a p_a^T} + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{U}\,\boldsymbol{Q^T} + \boldsymbol{F} = \sum_{a=1}^{A} \boldsymbol{u_a q_a^T} + \boldsymbol{F} \qquad \text{Equation 2.5}$$

Where **T** (*K x A*) and **U** (*K x A*) are the scores matrices for the blocks **X** and **Y**, respectively; matrices $\mathbf{P^T}$ (*A x N*) and $\mathbf{Q^T}$ (*A X P*) are the loadings matrices for blocks **X** and **Y**, respectively, and **E** and **F** are the residual matrices for block **X** and **Y**, respectively. [42], [46]

According to [44], for each factor, ***a*** = 1, 2, ..., *A*, to be included in the calibration process the followings steps are implemented:

(a) Calculate the loading weight vector, $\boldsymbol{w_A}$, by maximizing the covariance between the linear combination of $\boldsymbol{X_{A-1}}$ and $\mathbf{Y}_{\boldsymbol{A-1}}$ given that $\boldsymbol{w_A}^{\mathbf{T}}.\boldsymbol{w_A} = 1$;

(b) The factor scores, ***t***, are estimated by projecting $\boldsymbol{X_{A-1}}$ on $\boldsymbol{w_A}$;

(c) The loading vector $\boldsymbol{p_A}$ is determined by regressing $\boldsymbol{X_{A-1}}$ on $\boldsymbol{t_A}$ and similarly $\boldsymbol{q_A}$ by regressing $\boldsymbol{y_{A-1}}$ on $\boldsymbol{t_A}$;

(d) From $(\boldsymbol{X_{A-1}} - \boldsymbol{t_A}.\boldsymbol{w_A}^{\boldsymbol{T}})$ and $(\boldsymbol{Y_{A-1}} - \boldsymbol{t_A}.\boldsymbol{q_A}^{\boldsymbol{T}})$ new matrices $\boldsymbol{X_A}$ and $\boldsymbol{Y_A}$ are formed.

The optimum number of factors to include in the model is normally estimated by analyzing the validation procedure, which will be discussed further in this section. [44], [46]

The regression coefficients, **B**, can be estimated as a function of the loading weight matrix, **W**, and the loadings of **X** and **Y**, **P** and **Q**, respectively, as follows:

$$\boldsymbol{B} = \boldsymbol{W}(\boldsymbol{P}^T.\boldsymbol{W})^{-1}\boldsymbol{Q^T} \qquad \text{Equation 2.6}$$

Contrary to PCA, in PLS the loadings do not coincide fully with the direction of the maximum variation since loadings are corrected with the purpose of maximizing the predictive capacity of matrix **Y**.

In calibration, the regression matrix **B** is used to predict a sample without the need to resolve it into scores and loadings matrices. Therefore, if the experimental data (spectrum) from a given sample is defined by the vector $\boldsymbol{x_i}$, the predicted *y* variable for that sample can be calculated according to:

$$\hat{y} = x_A^T B \qquad \text{Equation 2.7}$$

In Figure 2.20, a schematic diagram of the PLS calibration and prediction processes is shown.

The PLS algorithm described above is used for multiple Y-variable and it is commonly known as PLS2. For single y-variable problems, the algorithm works in the same way but the ***y***-block is a vector and not a matrix, resulting in a simplified version of PLS2. [42]

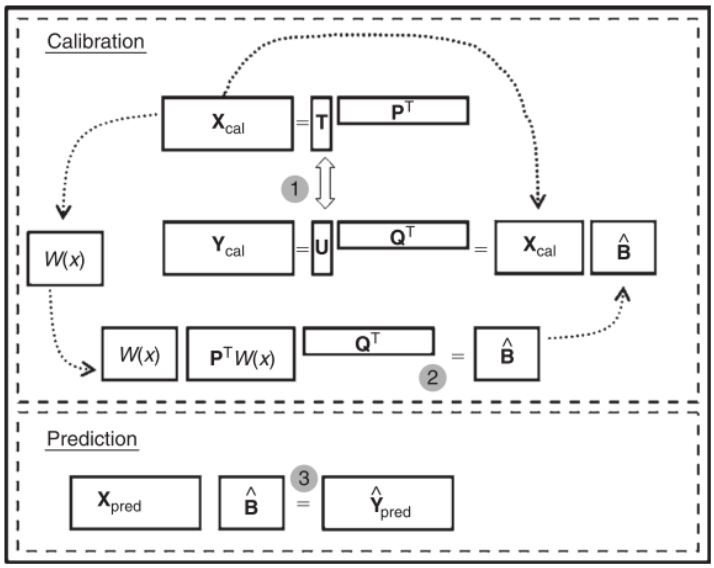


Figure 2.20 General schematic representation for the calibration of a PLS model and the prediction of new samples. Steps: (1) Simultaneous decomposition of ***X*** and ***Y*** matrices and calculation of weights matrix. (2) Calculation of regression matrix. (3) Prediction of new samples. [42]

MODEL VALIDATION

After the calibration of the regression model, it is of great importance to verify its performance as a multivariate method, being this related to modelling and interpretation, discrimination or prediction. This task is known as model validation and it is a very important step not only in terms of prediction performance but also in classification, since it can be used to avoid design the model to perfection on training samples at the cost of an inferior classification ability of new samples.

The validation process is essential in order to make sure that the regression model will work properly in future similar data sets, which can be seen as a prediction error estimation. Additionally, validation is also extremely useful to find the optimum number of latent variables for a model, by avoiding either overfitting or underfitting or incorrect interpretation. For some purposes, validation can also be applied in exploratory analysis methods, such as PCA.

When the purpose is to create a calibration model to predict quantities, such as concentration, it is possible to perform the model validation by using a test-set with an appropriated size. However, it

is not always possible to have reasonable amounts of objects as test-set due to the cost of the samples or referencing testing. As an alternative to the independent test-set validation, cross-validation (CV) is commonly applied.

Cross-validation consists of a validation method that uses the same objects for the model estimation and testing by leaving out a few objects out of the calibration data set and calibrating the model with only the remaining objects. Therefore, the objects left-out are predicted and the prediction residuals are computed. This process is repeated with other subsets of the calibration set until every object has been left out once. Then all the prediction residual are combined to calculate the validation residual variance and the root mean square error in prediction/cross-validation (RMSEP/RMSECV).

It is extremely important to use the adequate number of cross-validation segments, that is, use an appropriated level of cross-validation. For instance, when there are replicates for a sample, it must be certain that the same replicates are put together in the same segment.

The validation variance is the figure of merit in this process, together with the RMSECV, for regression models. It is normally referred that test-set validation is suitable for greater data sets (> 50 samples sets) and cross-validation is more adequate for small to medium data sets. The RMSECV is calculated according to the following expression:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad \text{Equation 2.8}$$

Where *n* is the number of objects/samples.

Depending on the level of validation chosen by the user, different schemes for cross-validation can be employed: Full CV - also known as leave-one-out – which leaves out only one object at a time; Segmented CV; Systematic segmented CV and CV across categorical information. For the purposes of the current work, apart from full CV, the remaining methods are not described. [42], [46]

Selection of the number of Latent Variables

When establishing a model, besides having the purpose of predicting or classifying new objects, there could also be the purpose of understanding the inherent structure of the system in study, which relates to the latent variables that convey the basic chemical or biological phenomena occurring in the system. Therefore, the interpretation of these models is highly dependent on its dimensionality, so the optimal number of latent variables to use must be investigated.

The number of latent variables can be selected in many ways, most of which relate to the prediction error obtained with the variable numbers. When a smaller number of optimal number of factors is used it causes the under fitting of the model and, thus, large errors. On the other hand, using a larger number of factors leads to overfitting, increasing the noise and causing larger errors.

The simplest method to select the optimum number of LVs is to plot RMSE (for calibration and validation – RMSECV for cross-validation or RMSEP for test-set validation) against the number of factors and choose the one corresponding to the minimum of the curve. This demands the assumption that the error decreases with increasing number of LVs until the point that the further factors contribute

mainly to noise and RMSE might increase due to the effect of overfitting. It is also possible that the error decreases continuously with the increase of the number of LVs making it difficult to verify the minimum error. Therefore, in order to solve this problem, another criterion was created, which involves selecting the number of factors with which the error is not significantly different from the minimum value for the model. This criterion can be applied by visual inspection of the plot for the RMSE vs the number of factors. The plot showing the explained variance for Y-variable against the number of factors can also be use, since the explained variance of *Y* and the RMSE are correlated with an *R* of -1 and, thus, this plot gives the same information as the RMSE plot. Figure 2.21 a) and b) show an example of the two plots that can be used to choose the number of LVs. [42], [46]

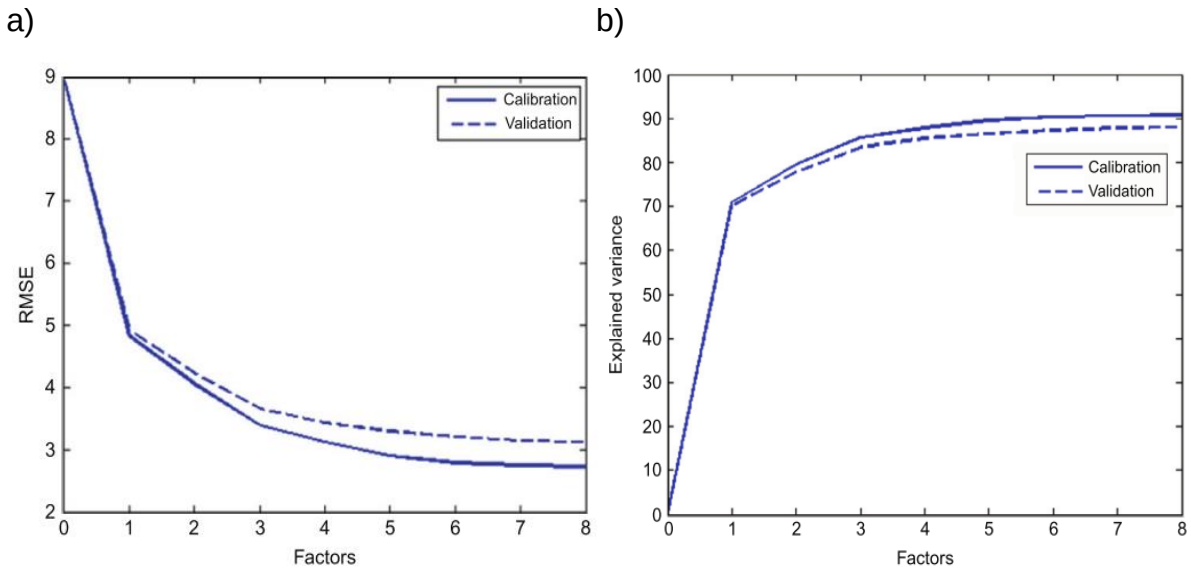


Figure 2.21 a) RMSE for calibration (solid) and validation (dashed) against the number of factors; b) Explained variance for one y-variable for calibration (solid) and validation (dashed) against the number of factors. [46]

Detection of Outliers

The same way as for PCA, in PLS regression is of extreme importance to verify the presence of outliers. The most commonly used plots to detect outliers are the plots of Leverage or Hotelling's T^2, which can be used to detect samples that are far from the center of the model space. Statistically, the critical limit of the Hotelling's T^2 and Leverage are based on a F-test and one *ad hoc* rule, respectively. However, as in PCA, the most commonly used plot to detect outliers is the influence plot, which displays the Hotelling's T^2 or the Leverage versus the *Q*-residual, as already described in section 2.4.1. [42], [46]

Pre-processing

The pre-processing techniques aim on removing the physical phenomena in the spectra with the objective of improving the subsequent multivariate data analysis – regression, classification model or exploratory analysis. [51] The pre-treatment of the data enables to suppress the effect on contributions which are not associated with the desired information in the spectra, in the interest of increasing the accuracy and precision.

Even though data pre-processing is a very important and extensive topic commonly used in spectral data, its various techniques will not be described due to the low necessity of using pre-

processing on the data used in the present work. For that purposes, only mean-centering and autoscaling is presented.

Mean-centering is a spectral scaling technique which relies on subtracting the average response from each individual response for the variable concerned, as follows:

$$X_{mc} = X - X_{avg} \qquad \text{Equation 2.9}$$

This means the variables are centered and, as a result, constant effects are suppressed.

Autoscaling consists on an initial centering of the data and subsequent division of each centered variable by the standard deviation for the responses of the variable concerned, as expressed bellow.

$$X_{mc} = \frac{X - X_{avg}}{X_{sd}} \qquad \text{Equation 2.10}$$

The autoscaled variables are centered in zero and have a unity standard deviation. [42]

2.4.3. Partial Least Squares Discriminant Analysis (PLS-DA)

The classification methods based on discriminant analysis rely on decision rules for the partition of the hyperspace of the variable in as many regions as the number of categories defined by the training objects. From a statistical point of view, these methods define the boundaries that separate the different classes in the multidimensional space using a criterion called Bayes' rule, which states that 'a sample should be assigned to the class to which it has the maximum probability to belong'. [52]

PLS-DA is a classification method based on the PLS algorithm, which was already described in section 2.4.2, but in a modified version to perform classification. The main differences relate to the dependent variables, as in PLS-DA these variables represent qualitative (and not quantitative) values.

In PLS-DA, the Y-block function as a dummy dependent matrix and it has as many columns as the number of classes and as many rows as the number of training objects. The information about the class of each sample from the training set is provided through a binary code: in the dependent matrix **Y** all entries of each row of the corresponding to a sample are set equal to 0, except for the column corresponding to the category the sample belongs to, whose element is set equal to 1. Therefore, as in PLS regression, **Y** is predicted. In this case, the predicted **Y** will not be compose of 1 and 0 values. Therefore, if there is only two classes a threshold is defined (for instance, 0.5), which decides if the sample is assigned to the corresponding class/column (calculated *Y* greater than 0.5) or not (calculated *Y* lower than 0.5).

When there is more than two classes, PLS-DA uses the algorithm PLS2, which refers to the PLS algorithm with several dependent *Y* variables. In this case, for each sample, the algorithm will return the prediction as a vector of size *G* containing values in-between 0 and 1: a value closer to 1 indicates that the corresponding sample belongs to the *i*th category, while a value closer to 0 represents the opposite. Since the values present in this vector are not the form (0, 0, …, 1, …,0), but rather values between 0 and 1, a classification rule is needed. There can either be by assigning to the class with the

maximum value in the predicted ***y*** vector or by imposing a threshold between zero and one appropriated for each category. [52], [53]

2.4.1. *K*-NEAREST NEIGHBORS (K-NN)

The k-NN is a non-parametric and distance-based technique and it is one of the simplest and oldest discriminant classifications methods. In fact, the classification of an unknown sample is performed by calculating its distance from each of the objects and detecting the *k* nearest objects (neighbors). Subsequently, the unknown sample is assigned to the group that has the most elements amongst these neighbors. In case of ties, the closer neighbors can acquire a greater weight.

The selection of the optimal number of neighbors is an important step in *k*-NN. This number influences the shape and smoothness of the defined boundaries that separate the different classes. The best selection procedure is to test a set of *k* values with a cross-validation approach, selecting the number of *k* that corresponds to the lowest classification error. However, in general, small numbers are used for *k*. When there are only two classes, odd numbers are used to prevent from tied votes, which makes 3 and 5 the most commonly used *k* values. When dealing with more than 2 classes a tie-breaking rule is required.

The success of k-NN can depend on the method used to measure the distance. Generally, the closeness of the samples is measured according to the Euclidean distance, which is expressed as follows:

$$d_{ui} = \sqrt{(x_u - x_i)^T (x_u - x_i)} \qquad \text{Equation 2.11}$$

Where $\mathbf{x}_u$ and $\mathbf{x}_i$ are the row vectors representing the coordinates of samples *u* and *i* in the multidimensional space. However, other distance measurement techniques, such as Mahalanobis distance, can be employed. It is also possible to perform the classification based on the scores from PCA, which works as a tool for the reduction of the data dimension.

This technique has the advantages of making no assumptions regarding the shape of the classes in the data space and handling with classification of multiple classes. Since the Euclidean distance is a non-linear function of the variables, another benefit of k-NN is being a non-linear classification method. [43], [52], [53]

2.4.2. SOFT INDEPENDENT MODELING BY CLASS ANALOGY (SIMCA)

Besides discriminant analysis, there is also another group of classification methods known as class-modelling methods, which is based on modelling peculiar features of the individual categories, focusing in its similarities rather than on what makes one class different from the others. One of the best-know class-modelling methods is SIMCA. [52]

SIMCA it is defined “soft” since there is no hypothesis on the distribution of variables and “independent” since the classes are modelled one at a time. This classification method is based on the exploratory analysis technique, PCA, coupled with the class information, by assuming that the relevant information about the similarities between the samples of each class is captured in a PCA model.

Therefore, SIMCA models consist of a compilation of *G* PCA models, one model for each of the *G* defined categories. Since the number of components from the training set might be different for each class, each PCA model is calculated separately on the objects of its class. Usually, cross-validation is used to choose the number of retained components for each class model.

Thus, a new object is projected in each of the G subspace (class model) defined by SIMCA and compared to it in order to verify its distance from the class. Hence, if the new samples are insufficiently close to the PC space of the class (through the analysis the Hotelling's T^2 plot, the *Q*-residual plot or the influence plot), then they are assigned as not belonging to this class. [40], [53]

3. MATERIALS AND METHODS

3.1. Overview

In order to achieve the purpose of this project an experimental procedure was conducted. This experiment can be divided into two main phases: pure spectra and screening phase, whose sample preparations are described in Section 3.2 and 3.3, respectively.

The samples used in this experiment were five compounds of *Carthamus* natural coloring product and four unwanted color additives - Orange II Sodium Salt, Tartrazine, Sunset Yellow and Annatto - of which the first three are synthetic dyes and the last one is a natural coloring product. The five *Carthamus* products were labelled from *A* to *E*, two of them (*B* and *C*) being the same product but from different batches. The details about these compounds can be seen in Table 3.1. All the samples were supplied by Chr. Hansen.

Even though Annatto is a natural product and therefore, is safe for human health, this compound in the context of this project will be treated as an unwanted compound since the addition of such compound in the *Carthamus* product jeopardizes its quality.

The spectroscopic methods/instruments used were: UV-Visible – since this is a simple method and Chr. Hansen has the instrument available; Fluorescence using two different measuring modes, front-face and right-angle - since the concentration levels of unwanted additives used are quite low and thus, a sensitive method like fluorescence seems appropriate - and, finally, Near-Infrared (NIR) as default method, commonly used in the analysis of food stuff. The sampling and measuring methods for the three instruments used are presented in Section 3.7.

Table 3.1 Details about the compounds used in this experimental procedure which were supplied by Chr. Hansen. B1 and B2 are the exact same product but were presented in different containers.

		Name	GIN	Batch	Label
***Carthamus* Natural Product**		Carthamus 30 CU	702986	0005068108	A
		Carthamus Liquid	708640	0005080030	B1/B2
		Carthamus Liquid	708640	0005078344	C
		Carthamus Liquid >7.5 CU	-	0005082816	D
		Natural Sweet FL 30	-	0005083695	E
Unwanted Additives	Synthetic	Orange II Sodium Salt	195235-25G Dye content > 85% Sigma Aldrich		
		Tartrazine	86310 Luka Chemika		
		Sunset Yellow FCF	465224-25 Dye content: 90% Sigma Aldrich		
	Natural	Annatto	A-720-WS-AP GIN: 242531 Batch: 3136680		

For the pure spectra phase, all the samples - *Carthamus* and all the unwanted dyes - were measured by three different spectroscopic methods: UV-Visible, Fluorescence (front-face and right-angle mode) and NIR. The aim of this part of the experiment is to test how the detection works for each sample and to determine the best settings for the spectroscopic method in use. To obtain the pure spectra no mixture was made and the compounds were measured in pure form or diluted in water according to the features of the instrument in use. For this phase, some of the samples were previously prepared as described on the experimental procedure presented in Section 3.2.

For the screening phase, the four color additives (Orange II, Tartrazine, Sunset Yellow and Annatto) were diluted into the natural product *Carthamus* (labelled as *B*) in different concentration levels: 0.8, 1.6, 8, 40, 80, 160 mg of pure additive in kg of *Carthamus* color product (or 1, 2, 10, 50, 100 and 200 ppm). These concentrations were chosen according to the information provided by Chr. Hansen, resulting from previous occurrences of contaminations in this product that registered concentrations in this range.

The main goal of the dilution made in the screening phase is to simulate different quantities of unwanted additives in the *Carthamus* samples with the objective of creating a model/method that could quantify these unwanted agents. In order to do this, all the samples prepared for the screening were measured as pure and/or as a diluted form (in a 1:1000 and 1:10000 proportion) in the same instruments used in the pure spectra trial. As for the pure spectra phase, in the screening, some of the samples were previously prepared according to the experimental procedure presented in Section 3.2.

Additional experiments - linearity and detection tests (Section 3.4) and optimization of the parameters/settings of the instruments (Section 3.5) – were performed, the results of which were used to improve the performance of the measurements made in the pure spectra and screening phase. Additionally, an experiment was conducted using the five *Carthamus* products which aimed at obtaining additional data on the pure *Carthamus* samples (Section 3.6).

The data obtained was further analyzed using chemometric methods (Section 3.8).

3.2. Sample preparation for pure spectra phase

Purpose

The sample preparation for the pure spectra phase was necessary only for the UV-Visible and Fluorescence. These samples will further be used to obtain the pure spectra in the referred spectroscopic methods.

Material and Equipment

Volumetric flasks of 50 mL
Analytical balance Mettler AJ100 METTLER TOLEDO
Micropipette BioHIT m100/m100

Experimental Procedure

- The synthetic compounds Orange II, Tartrazine and Sunset Yellow, presented in powder, were weighed and transfer to a volumetric flask of 50 mL and the five different *Carthamus* compounds and the natural coloring compound Annatto, as liquids, were pipette also to volumetric flasks of 50 mL.
- All the compounds were diluted in water to achieve the adequate dilution ratio.
- The volumetric flasks were then wrapped in foil to protect from sun light and stored in a dark room refrigerated at an average temperature of 4ºC, until further measurements.

3.3. Sample preparation for screening phase

For the screening phase a set of different mixtures were prepared. The compounds used in this solutions are presented in Table 3.1. The *Carthamus* used was the one labelled as *B*.

Purpose

The purpose of this preparation is to make six mixtures of pure *Carthamus* with each of the four unwanted additives in different concentrations. Initially, a set of four stock solutions were prepared for each unwanted agent and then these stock solutions were used to prepare the remaining necessary solutions.

Materials and Equipment

Volumetric flasks of 50 mL
Beakers of 50 mL
Glass rod
Blue cap flasks
Graduated cylinder
Micropipette BioHIT m100/m100
Analytical balance Mettler AJ100 METTLER TOLEDO

Experimental Procedure

- In order to prepare the stock solutions, the four color additives (Orange II, Tartrazine, Sunset Yellow and Annatto) were diluted in *Carthamus B* in the greatest concentration previously established (160 mg of pure additive in kg of *Carthamus* color product). This was made using a micropipette for the dye Annatto (liquid form) and the non-natural color additives, as powders, were weighed and then added to *Carthamus* with a spatula.

- The homogenization of the unwanted additives and the *Carthamus* was made manually in a beaker using a glass rod, being then transfer to volumetric flasks where the remaining amount of *Carthamus* was added.

- In order to prepare the remaining mixtures with the concentration previously established:0.8, 1.6, 8, 40, 80 mg of pure additive in kg of *Carthamus* color product, a precise volumetric quantity of *Carthamus* pure was added to a precise amount of the stock solution previously prepared.

- All the 24 different mixtures prepared were wrapped in foil to protect from sun light and stored in a dark room refrigerated at an average temperature of 4ºC, until further measurements.

3.4. Sample preparation for linearity and detection tests

Purpose

The main objective of the linearity test is to understand if there is linearity in the concentration of coloring agents contained in a certain solvent, in this trial the solvent used is water. On the other hand, the detection test has the purpose of verifying if the instrument is detecting the presence of these unwanted coloring agents in water, using the same range of concentrations as the ones used with the *Carthamus* in pure form and after dilution in water.

In order to perform the linearity and detection test for the UV-Visible spectrophotometer, 21 different samples were prepared, 7 for each of the synthetic additives – Orange II, Tartrazine and Sunset Yellow. For the detection test made for the fluorescence spectrophotometer, 3 different samples were prepared, for each synthetic color additive, since only one level of concentration was tested (160 ppb).

Materials and Equipment

Volumetric flasks of 100 mL
Analytical balance Mettler AJ100 METTLER TOLEDO
Micropipette BioHIT m100/m100

Experimental Procedure

- Weight the previously defined amount of color additive and transfer it to the corresponding volumetric flask of 100 mL. Repeat for each synthetic color additive.
- Fill the rest of each volumetric flask with water.
- Homogenize the solutions, wrap the volumetric flasks in foil to protect from sun light and store until further measurements.

3.5. Sample preparation for the optimization of the parameters involved in the fluorescence spectrophotometer

Purpose

The main purpose of this procedure is to prepare samples with different dilutions in water (1:10 000; 1:1000; 1:100; 1:10) that would be used to understand which dilution ratio is more appropriated and gives a better signal resolution in the fluorescence spectrophotometer. Furthermore, the samples prepared in this phase are also necessary to optimize the settings and setups used in the fluorescence instrument in front-face mode.

Materials and Equipment

Micropipette BioHIT m100/m100
Volumetric flasks of 100 mL

Experimental Procedure

- Choose a random mixture prepared for the screening phase and transfer, using a micropipette, the 4 previously calculated quantities of mixture to 4 volumetric flasks in order to achieve the 4 predetermined levels of concentrations required (1:10 000; 1:1000; 1:100; 1:10).
- Fill the rest of each volumetric flask with water.
- Homogenize the solutions, wrap the volumetric flasks in foil to protect from sun light and store until further measurements.

Optimized settings

After performing the experimental procedure described above, it was concluded that the optimized settings were the ones shown in Table 3.2. Additionally, it was observed that the optimum dilution ratio is 1:1000.

Table 3.2 Optimized settings for the fluorescence spectrophotometer (a new setup for the sample holder of the instrument was also used).

	Start	Step	End	Slit
Excitation (nm)	225	5	450	5
Emission (nm)	220	2	620	5
Scan control	Manual			
Scan rate	4800 nm/min			
Detector voltage	Manual - 850 V			

3.6. Preparation of additional pure *Carthamus* samples

Purpose

The aim of this experiment is to increase the amount of data on the pure *Carthamus* products when measuring in UV-Visible spectrophotometer, since only 5 *Carthamus* products were available. In order to do this, a set of mixture of the five products of *Carthamus* in different proportions was made. A set of 22 different mixtures were planned, containing the five different *Carthamus* and using three different proportions: 50:50, 25:75 and 75:25.

Materials and Equipment

Volumetric flasks of 50 mL
Micropipette BioHIT m100/m100

Experimental Procedure

- Transfer the predefined quantity of the two different *Carthamus* which form the desired mixture with a micropipette to a volumetric flask in order to achieve the right dilution proportion with water.
- Fill the rest of the volumetric flask with water.
- Homogenize the solution, wrap the volumetric flask in foil to protect from sun light and store until further measurements.

3.7. Spectroscopic measurements and sampling

Purpose

In order to measure all the different samples, four different spectroscopic methods were used: UV-Visible, NIR and Fluorescence in front-face mode and in right-angle mode.

All the settings used in each instrument for each type of sample are presented in Appendix I.

UV-VISIBLE

Material

3 mL *Pasteur* pipette FRISENETTE
Disposable cuvettes VWR

Equipment

Thermo Scientific Evolution 220
Software: Thermo Scientific Insight

Experimental Procedure

- A blank background was measured without any sample.
- The samples were transfer with a pipette from the flask to a disposable UV-cuvette.
- Each sample was measured three times. In between each replicate the sample was transfer back to the flask, and after homogenization of the solution, transfer back again to the cuvette to be measured.
- A new disposable cuvette was used for each sample.

FLUORESCENCE

Material

Cuvette QS High Precision Cell 10x10mm HELLMAANALYTICS

3 mL *Pasteur* pipette FRISENETTE

Equipment

Air Gun CEJN

VARIAN Cary Eclipse

Software: Cary Eclipse Scan application

In this instrument, an in-house code (built by the SPECC group) was used to measure the samples instead of using the predefined functions of the Cary Eclipse software. This code functioned in such a way that the light with the lowest energy (highest wavelength) was measured first in order to minimize the probability of the occurrence of any chemical reaction taking place due to the added energy to the sample.

Experimental Procedure

For liquid compounds (Front-face and right-angle mode)

- For front-face and right-angle mode the appropriated configuration was assembled, as can be seen in Figure 3.1 and Figure 3.2, respectively. In the first figure, the red arrows demonstrate how the light travels in the instrument.
- The solutions were transfer from the containers to the cuvette with a *Pasteur* pipette.
- Each sample was measured three times. In between each replicate, the sample was transfer back to the flask and, after homogenization of the solution, transfer again to the cuvette to be measured. All the sample were measured in an EEM (landscape).
- Between each sample the cuvette was washed with water and dried with an air gun.

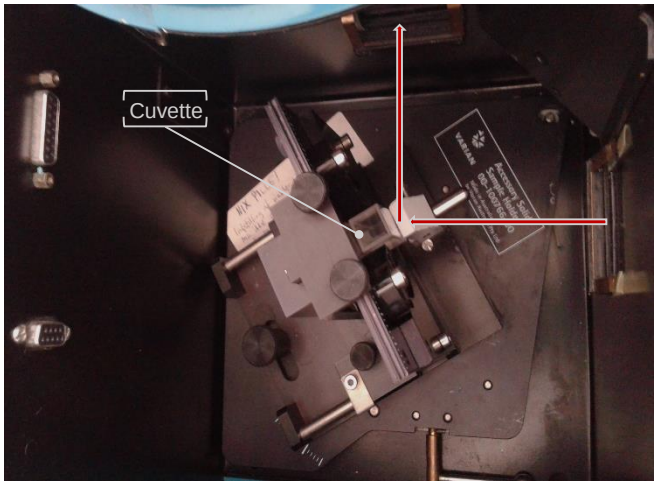


Figure 3.1 Configuration for measuring liquids in a cuvette through front-face mode. The red arrows show how the light travels from the right to the sample and then up to the detector.

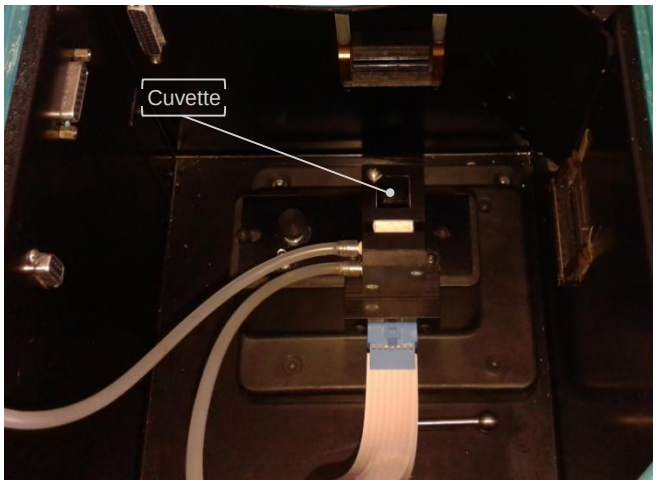


Figure 3.2 Configuration for measuring liquids in a cuvette through right-angle mode.

For solid compounds:

- The instrument assembling part for solids was installed, as can be seen in Figure 3.3. In this figure, the red arrows demonstrate how the light travels in the instrument.
- Using a spatula a small amount of the compound was transferred to the solid sample holder;
- The sample was measured in landscape.
- Between each sample the solid sample holder was cleaned with a dry tissue.

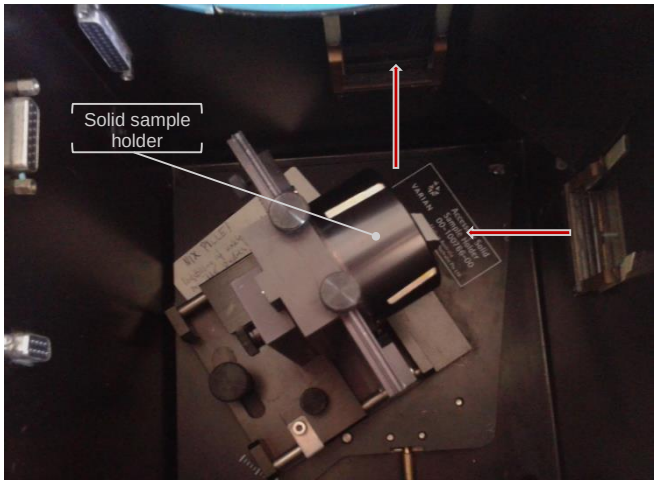


Figure 3.3 Solid sample holder configuration (front-face mode). The red arrows show how the light travels from the right to the sample and then up to the detector.

NEAR -INFRARED

Material

20 mL disposable scintillation vials
Reference: PTFE Standard

Equipment

QFA flex Q-interline
Spinning module
Software: Grams/LT Version 7

Experimental Procedure

- A background using a PTFE standard was made;
- The sample (liquid or solid, as pure form) were transferred to the scintillation vials;
- Each liquid sample was measured three times and, in between each replicate, the sample was transfer back to the container and, after homogenization, transfer again to the scintillation vial to be measured.
- Each scintillation vial was placed in the spinning module and sampled.

3.8. Data Analysis

The data acquired from all the spectroscopic methods used were exported from each instrument software and converted to Matlab files in MATLAB Version 2013a (the Matlab scripts used for this task are shown in Appendix II).

Subsequently, the Chemometric analysis was performed in LatentiX 2.12 (more information about this multivariate analysis software here: [54]). Some additional calculations were executed in Microsoft Office Excel 2013 and in MATLAB.

4. RESULTS AND DISCUSSION

4.1. Overview

In this section all the results obtained from the work performed in the laboratory and, subsequently, from the multivariate analysis are presented and discussed.

Initially, the raw spectra acquired from the three spectroscopic instruments for the pure spectra and screening phase are displayed in Section 4.2 and 4.3, respectively, using LatentiX for its visualization.

Subsequently, the results for the multivariate analysis on the spectral data acquired are shown, analyzed and discussed. The multivariate analysis is divided into three main parts: exploratory analysis, quantitative analysis and qualitative analysis.

The exploratory analysis was performed, using PCA, in the spectral data for the three spectroscopic methods – UV-Visible, Fluorescence and NIR. After analyzing the PCA results, it was considered that the UV-Visible and Fluorescence data showed the best data structure to be used in the quantification and identification of the unwanted additives in *Carthamus.* Although the UV-Visible showed slightly superior results, the Fluorescence data showed promising results, given that the data resolution might improve if the measurements parameters (concerning the setup and settings of the instrument and the sample preparation) were optimized. On the other hand, the NIR data showed significantly inferior results and, therefore, it was considered as not useful for further analysis.

In the first instance, a quantitative analysis was implemented in the UV-Vis data, through PLS regression. Consecutively, the qualitative analysis was performed in the same data using two distinct classification methods – PLS-DA and SIMCA.

Since the results obtained with the UV-Vis data were not entirely satisfactory, the detection ability of the instrument and the linearity of the data were tested. The results of these tests showed that there is linearity in the concentration of the unwanted additives, however, the detection capacity of the UV-Vis instrument, for such low quantity of additives, is quite weak.

As the results previously obtained did not fully comply with the aim of the present study, the fluorescence data was further analyzed. Firstly, the parameters related with the measurements made in the spectrofluorometer were optimized to improve the resolution of the data that had been already acquired. Afterwards, the new data acquired was investigated and quantitative and qualitative analysis were performed. As in for UV-Vis, a quantitative analysis was performed, using PLS regression. For the qualitative analysis the classification methods used were PLS-DA and k-NN. The latter was used because it was verified a non-linear trend in this data, so k-NN would be more appropriated than PLS-DA. Additionally, a detection test was also executed for the fluorescence instrument, as previously done for UV-Visible.

4.2. Pure Spectra

The pure spectra was obtained by measuring the 9 compounds supplied by Chr. Hansen in pure form or diluted in water (no mixture was made in this phase) using UV-Visible, Fluorescence and NIR. The 9 compounds consist of 5 natural coloring *Carthamus* products, with two of them (*B* and *C*) being the same product but from different batches, and 4 unwanted color additives, three of them being synthetic – Orange II, Tartrazine and Sunset Yellow - and one being from natural origin - Annatto.

After obtaining the data from the instrument software and further conversion to Matlab file, the data was analyzed using LatentiX.

UV-VISIBLE

For the UV-Visible spectroscopy, the samples required a dilution as *Carthamus* is a very thick and opaque liquid that cannot be measured by the instrument in pure form, as the sample will absorb all incoming light. Considering that this product is soluble in water, this was the solvent chosen. Thus, the first step was to perform a dilution test, in which different dilution rates were used. By analyzing the response from the instrument detector to the different samples prepared, and considering that its saturation must be avoided, an appropriated dilution rate was chosen for each sample.

In Figure 4.1 the pure spectra for all the different *Carthamus* obtained with UV-Vis spectrophotometer is shown. As can be seen, the five compounds of *Carthamus* have the same tendency but the intensity of the peaks is quite different, which has to do with their color strength. The compounds *Carthamus A* and *E* have higher color strengths, *Carthamus D* has the weakest color strength and *B* and *C* are in between these two levels of color. Even though *C* and *B* are not exactly overlapping each other, they are fairly similar, which was expected since they are the same product but from different batches.

By analyzing all the curves represented in Figure 4.1, it is possible to understand that either *B* or *C* are appropriated candidates to use in further analysis since they have a medium color strength which is useful to create more comprehensive models in future analysis and, at the same time, a big amount of these two products was supplied by Chr. Hansen (*Carthamus C*). From these two products, the *Carthamus* B was the one chosen to be used in further analysis but for no particular reason as both, B and C, shared the same features.

It can also be seen that the region that seems more interesting in this spectra is from 300 to 550 nm since it shows the two most relevant peaks.

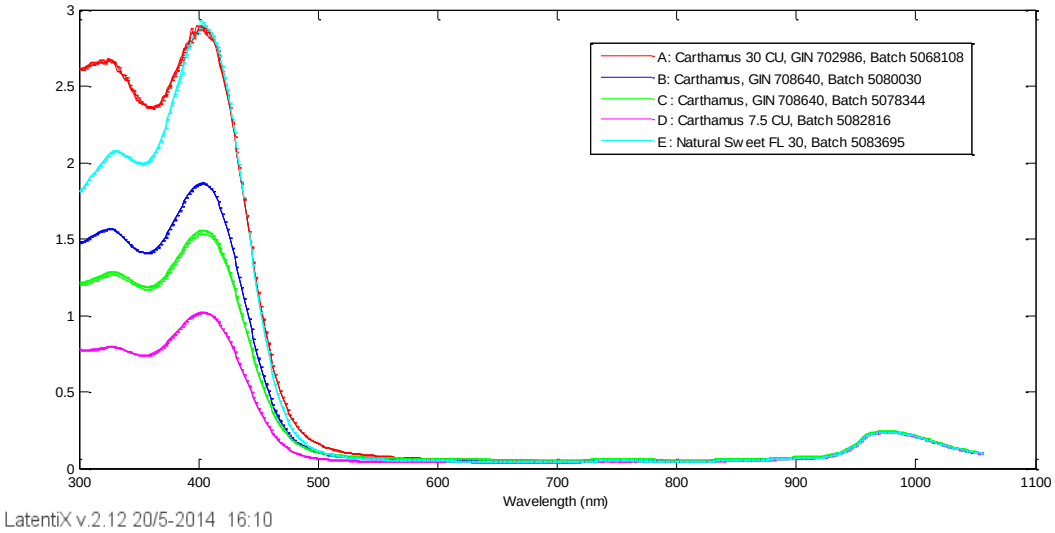


Figure 4.1 UV-Visible pure spectra for the different natural coloring product Carthamus. These samples were diluted in water to reach a 1000 ppm solution.

In Figure 4.2, the pure spectra for the four unwanted coloring agents are presented. In order to measure these additives a dilution test was also performed with the purpose of understanding which dilution ratio was suitable for each additive.

Therefore, the synthetic dyes - Orange II, Tartrazine and Sunset Yellow - were diluted in water in order to achieve a 16 ppm solution, whereas the natural dye - Annatto - was diluted in water to reach a 1000 ppm solution. The dilution ratio used was not the same for all of them since Annatto had been previously diluted by Chr. Hansen and it showed a weaker signal comparing to the remaining additives and, in this sense, a further dilution of this compound would turn this signal imperceptible in the plot. In contrary, if a lower dilution was used for the synthetic additives, the detector would get saturated.

In this figure, the region of interest showing the more relevant peaks is from 280 nm to 600 nm. The first peak is present only for Orange II and Sunset Yellow and, although Orange II shows a more intense peak around 300 nm, the overall spectra for these two compounds have a similar trend. The spectrum for Tartrazine is very smooth displaying only one peak around 400 nm. Annatto shows a low intensity peak around 350 nm, not present for the remaining additives, possibly because this is a natural compound and, thus, different from the other synthetic compounds. However, in general, all the unwanted coloring agents display a lot of information in the spectral range between 350 nm and 550 nm.

By briefly analyzing Figure 4.1 and Figure 4.2, it can be observed that in the spectral range from 320 to 420 nm, in the first figure, *Carthamus* substances show a substantial signal, contrary to what can be observed in Figure 4.2, in which the unwanted additives do not show a significant peak. Also, in the region between 470 and 600 nm, the unwanted additives still have signal, especially Orange II and Sunset Yellow, but for the *Carthamus* products there is no signal anymore. The two regions mentioned above (320-420 nm and 470-600 nm) might be useful and valuable to distinguish the pure *Carthamus* from the impure in further analysis.

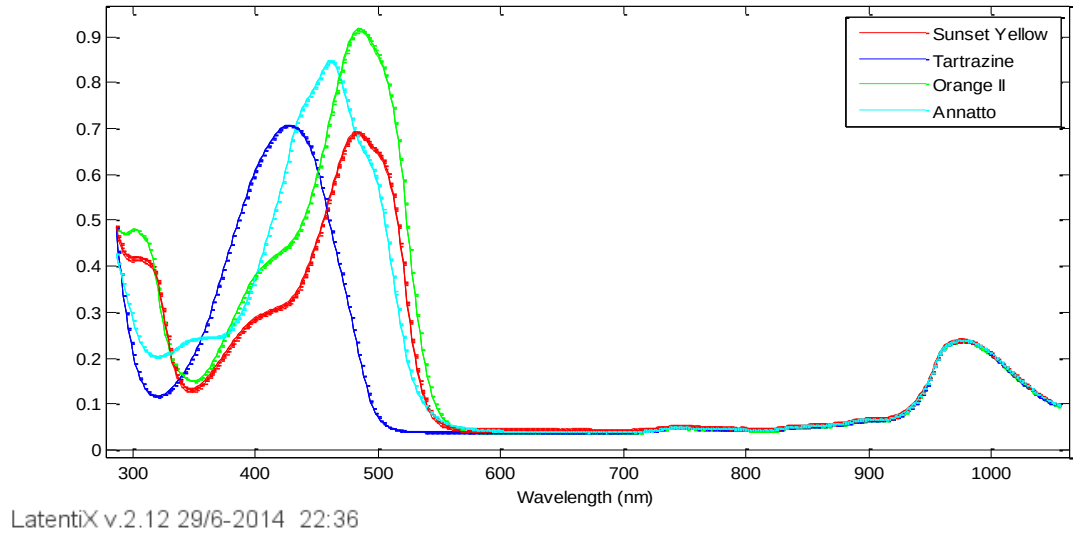


Figure 4.2 UV-Visible pure spectra for the four unwanted agents. The synthetic dyes - Orange II, Tartrazine and Sunset Yellow - were diluted in water in order to achieve a 16 ppm water solution, whereas the natural dye - Annatto - was also diluted in water to reach a 1000 ppm solution.

FLUORESCENCE

Figure 4.3 and Figure 4.5 represent the pure spectra for *Carthamus* substances obtained in fluorescence using front-face and right-angle mode, respectively. Since the measurement was made using an Excitation and Emission Matrix (EEM) and the three-dimensional plots are complex figures, appropriated excitation regions were chosen to present a two-dimensional plot. The first figure, for the front-face mode, shows the same information as the results acquired in UV-Vis, in which there is a separation between three groups of *Carthamus* - A and E, B and C and D – according to the color strength.

In Figure 4.4 and Figure 4.6 the pure spectra for the unwanted additives is demonstrated for both excitation regions, 340 nm and 290 nm, respectively, in order to properly compare them with the *Carthamus* pure spectra.

By comparing Figure 4.3 and Figure 4.4, for front-face mode, it is possible to notice that in the emission region from 400 to 500 nm there is very small signal for the pure Carthamus samples, however the unwanted additives show a strong signal, which indicates that this spectral range might be interesting and useful to distinguish pure *Carthamus* from the *Carthamus* contaminated with the unwanted additives in further analysis.

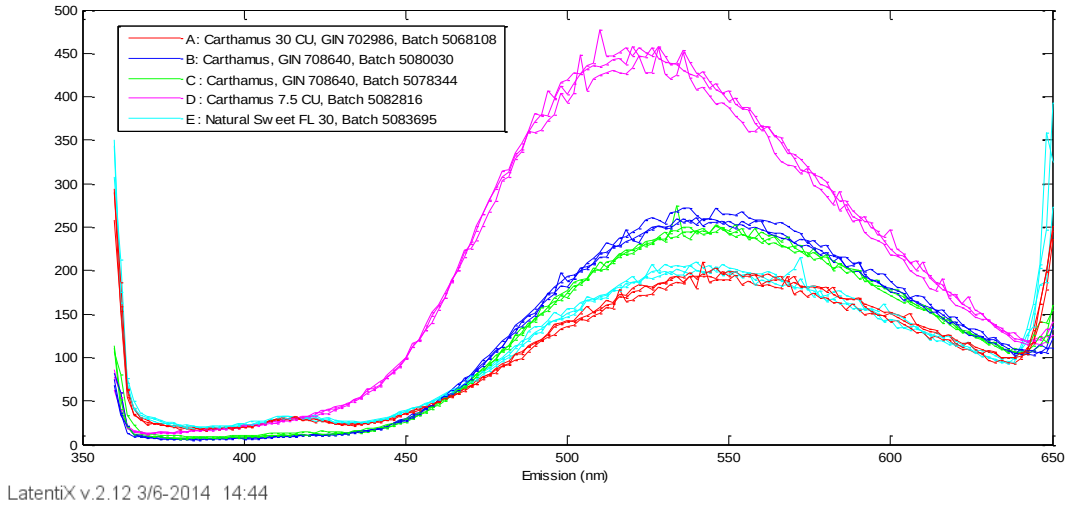


Figure 4.3 Fluorescence pure spectra for the different natural coloring product Carthamus in pure form (Excitation region: 340 nm; Front-face mode).

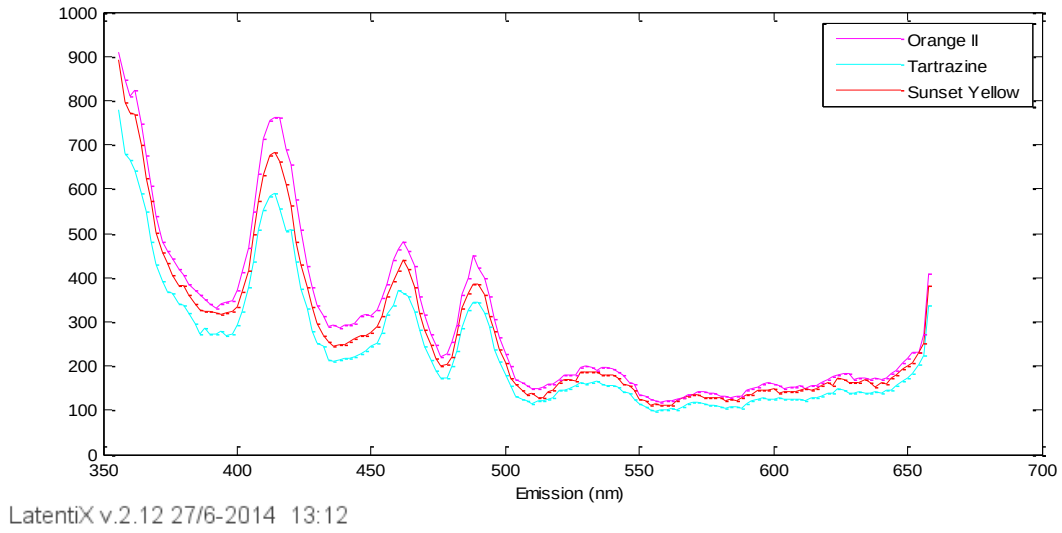


Figure 4.4 Fluorescence pure spectra for the three different synthetic agents (Excitation region: 340 nm; Front-face mode).

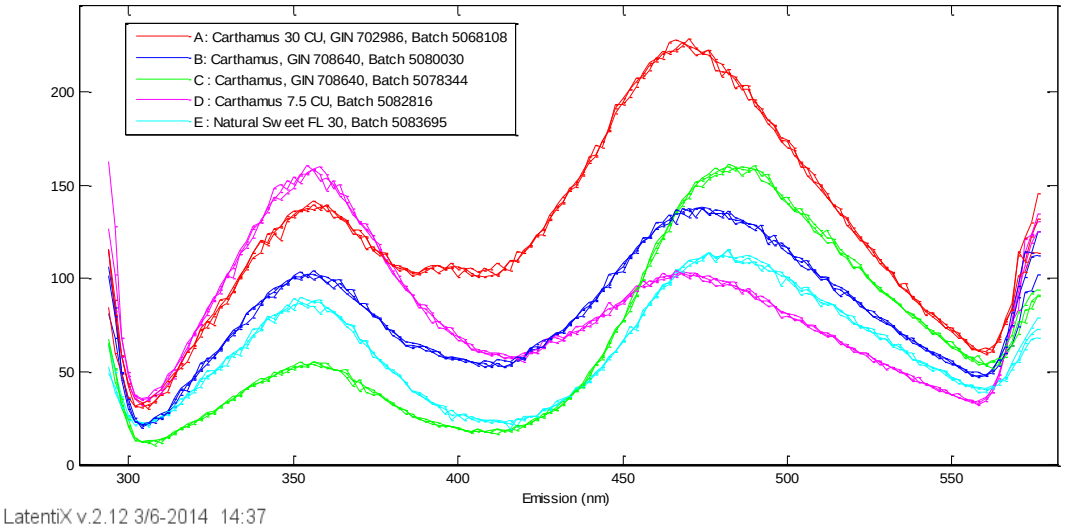


Figure 4.5 Fluorescence pure spectra for the different natural coloring product Carthamus diluted in water (Excitation region: 290 nm; Right-angle mode).

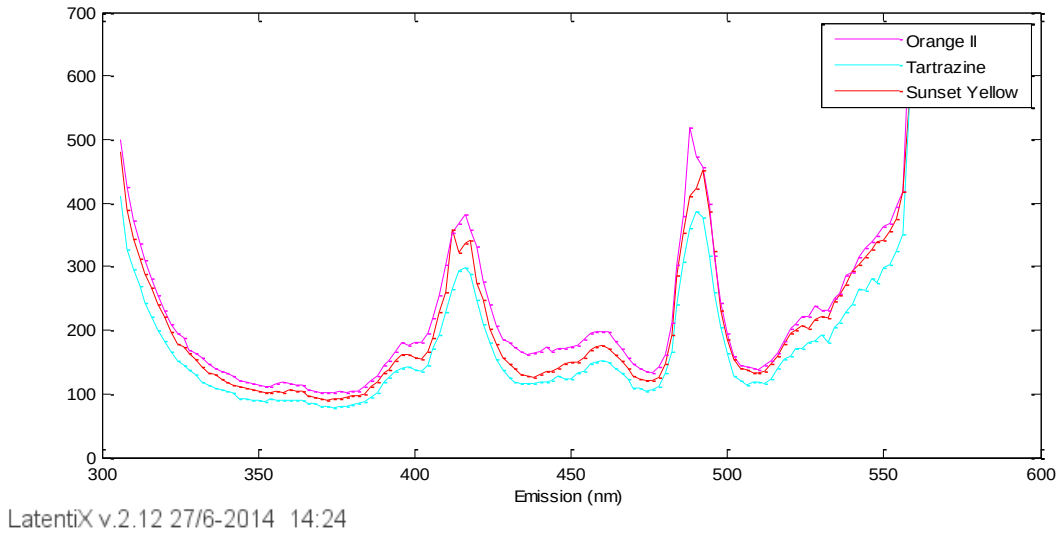


Figure 4.6 Fluorescence pure spectra for the three different synthetic agents (Excitation region: 290 nm; Front-face mode).

NEAR-INFRARED

In Figure 4.7, the pure spectra for all the Carthamus compounds through NIR spectroscopy are shown. In this case, *Carthamus* E is quite different from the others and the remaining compounds are very similar. And, in that sense, it is not possible to uptake a lot of information from this figure.

Figure 4.8 Figure 4.9 show the spectra for the unwanted agents. The first figure shows that in the region from 7500 to 5000 cm^{-1} (~1333 to 2000 nm) Sunset Yellow and Tartrazine are quite similar, however in the peak at 6000 cm^{-1} Orange II and Sunset Yellow have the same behavior. Since the natural additive, Annatto, is presented in liquid form and the others color additives are powders, it is not possible to analyze these together.

It is not possible to compare the pure *Carthamus* samples and the synthetic additives, since *Carthamus* are presented in liquid form and the synthetic additives as powders.

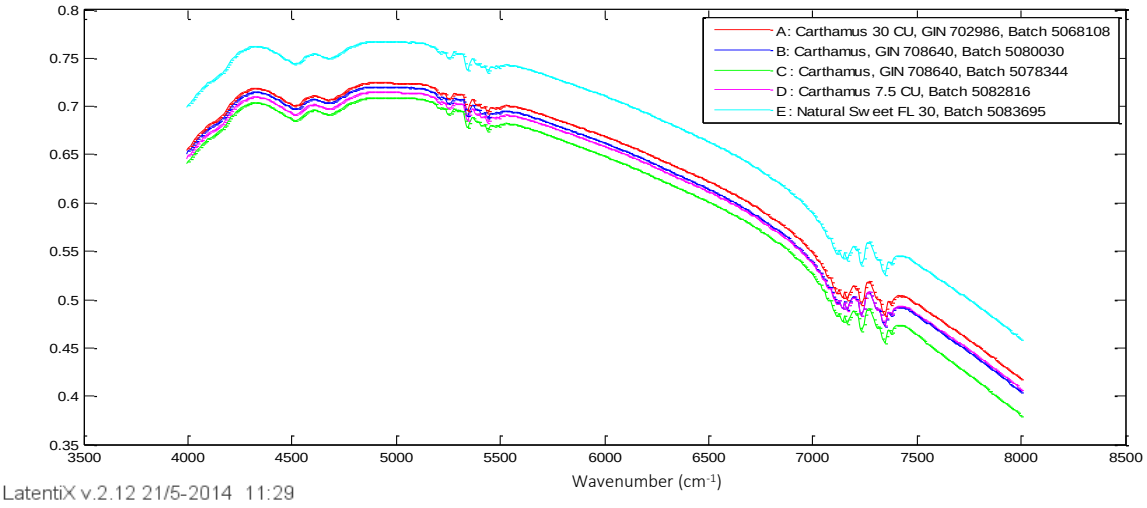


Figure 4.7 NIR pure spectra for the different natural coloring product Carthamus (Resolution: 8 cm-1; 64 scans).

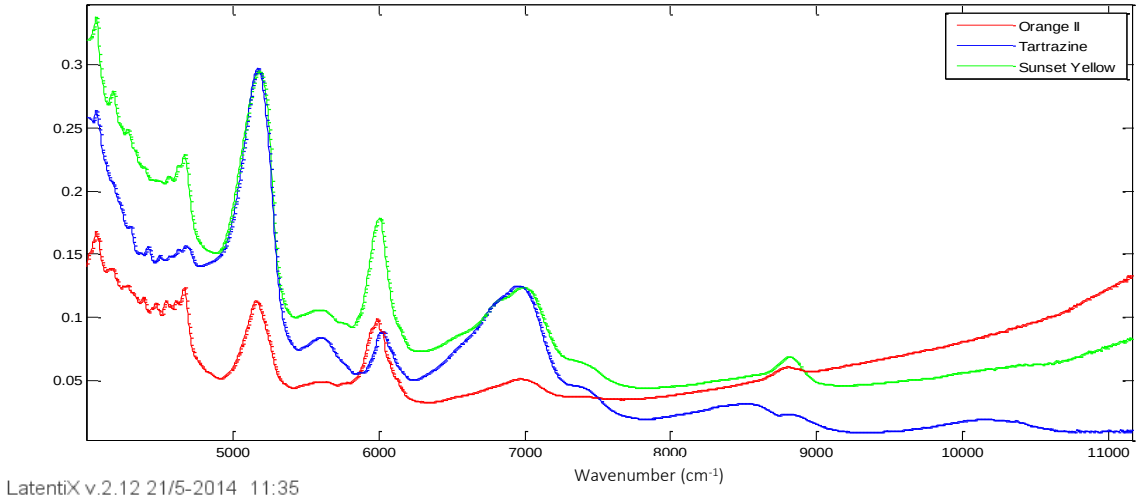


Figure 4.8 NIR pure spectra for the pure synthetic dye in powder form (Resolution: 8 cm-1; 64 scans).

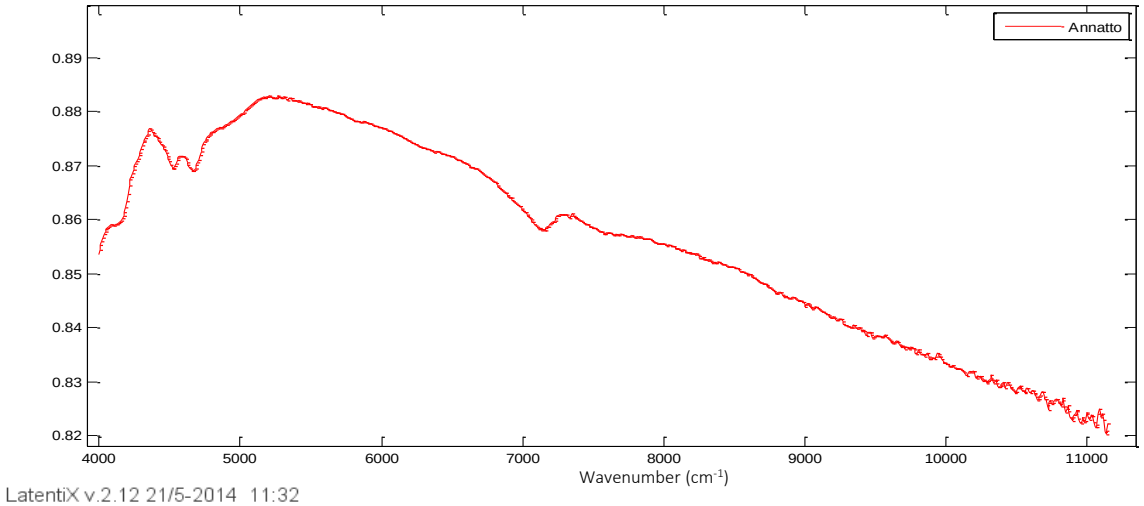


Figure 4.9 NIR pure spectra for the liquid natural color, Annatto. This one cannot be shown together with the remaining color compounds since Annatto is in a liquid form and the remaining are powders (Resolution: 8 cm^{-1}; 64 scans).

4.3. Screening Spectra

For the screening phase the four unwanted additives were diluted in the *Carthamus B* in six different concentrations: 0.8, 1.6, 8, 40, 80, 160 mg of pure additive in kg of *Carthamus* color product (or 1, 2, 10, 50, 100 and 200 ppm), giving a total of 24 different mixtures. The main purpose of this screening phase was to acquire data that can be used to create a quantitative model that could quantify the unwanted additives present in *Carthamus* and, subsequently, develop a classification model able to distinguish the pure *Carthamus* samples from the impure ones and possibly detect which coloring agent is present in each impure sample. Therefore, the 24 samples prepared and the 5 pure *Carthamus* were measured using three spectroscopic instruments and the resulting spectra are presented below.

UV-VISIBLE

For the UV-Visible the 29 samples were diluted in water in the same concentration as for the pure spectra for the pure *Carthamus* samples (1000 ppm). Figure 4.10 a) shows the resulting spectra for all the samples measured in the screening trial for UV-Visible. To better visualize the differences between the pure and the impure *Carthamus* samples, Figure 4.10 b) only displays the spectra for the pure *Carthamus* samples and the highest concentration for each of the unwanted agents added to *Carthamus.*

From these two figures it can be observed that when the unwanted agents are added to the pure *Carthamus* only the intensity of the peak is modified and there are no changes in the shape. This behavior is quite peculiar since the pure spectra for the unwanted additives showed different trends comparing with pure *Carthamus*.

Even though these spectra show a strange behavior, a distinction between the spectra for pure and impure samples is observed in the figures below, which means that a separation between the pure and impure *Carthamus* samples is possible and, for that reason, it would be interesting to use this data in further analysis.

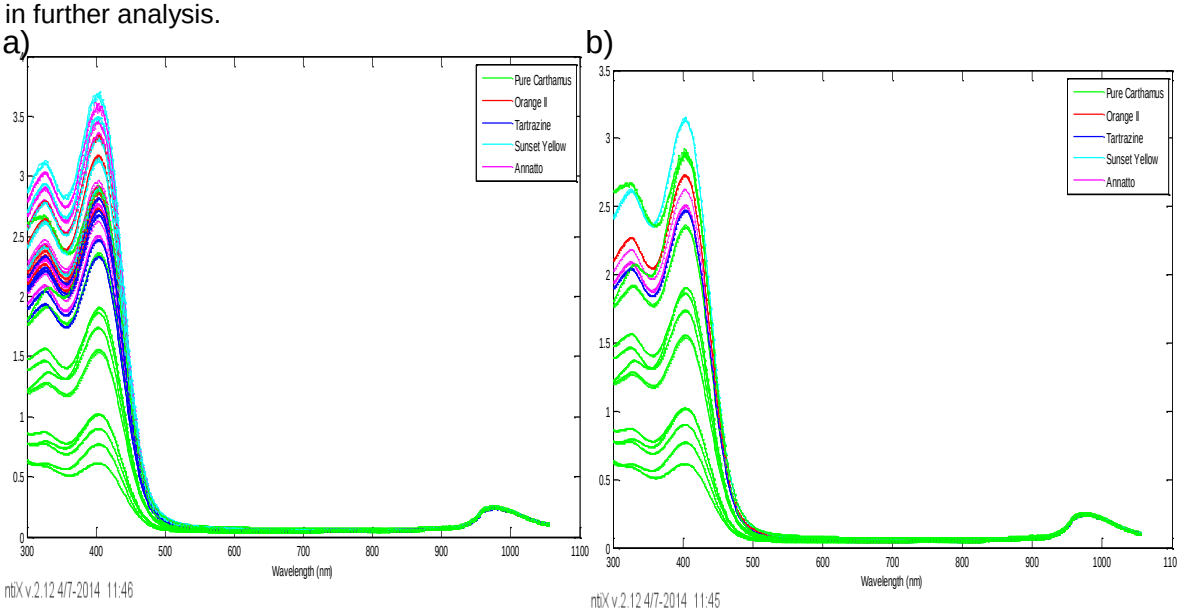


Figure 4.10 Raw spectra for the screening phase using UV-Visible spectroscopy; a) showing all the pure and impure samples; b) showing only the pure Carthamus samples and the ones in which the highest quantity of unwanted color additive had been added (160 mg of pure additive in kg of pure *Carthamus*).

FLUORESCENCE

In this section the raw spectra for fluorescence screening phase is shown and analyzed.

Figure 4.11 a) presents the spectra for the samples measured in front-face mode. This plot shows that there is no clear distinction between the pure samples and the adulterated samples, however, in the emission region from 370 nm to 450 nm there is a signal for the adulterated *Carthamus* samples, while the pure samples do not seem to show any information in that region, which indicates that this region might be interesting to separate the two categories, pure and impure *Carthamus*.

In Figure 4.11 b) the samples measured in right-angle mode are shown. In this case, the samples were diluted in water in order to achieve a 1000 ppm solution, just as has been done for the samples used in UV-Visible. By analyzing this figure it can be seen a distinction between the pure samples and the remaining ones containing the unwanted additives, even though is not a completely clear separation. Moreover, the samples containing the four different unwanted additives are indistinct from each other which indicate that it would be difficult to identify which coloring additive is present in *Carthamus*.

At first glance, the data obtained in fluorescence, either using front-face or right-angle mode, does not seem as valuable as the UV-Visible data. However, it could be interesting to further analyze this data in order to understand if any adjustment to the instrument; such as setup and the settings used, should be done.

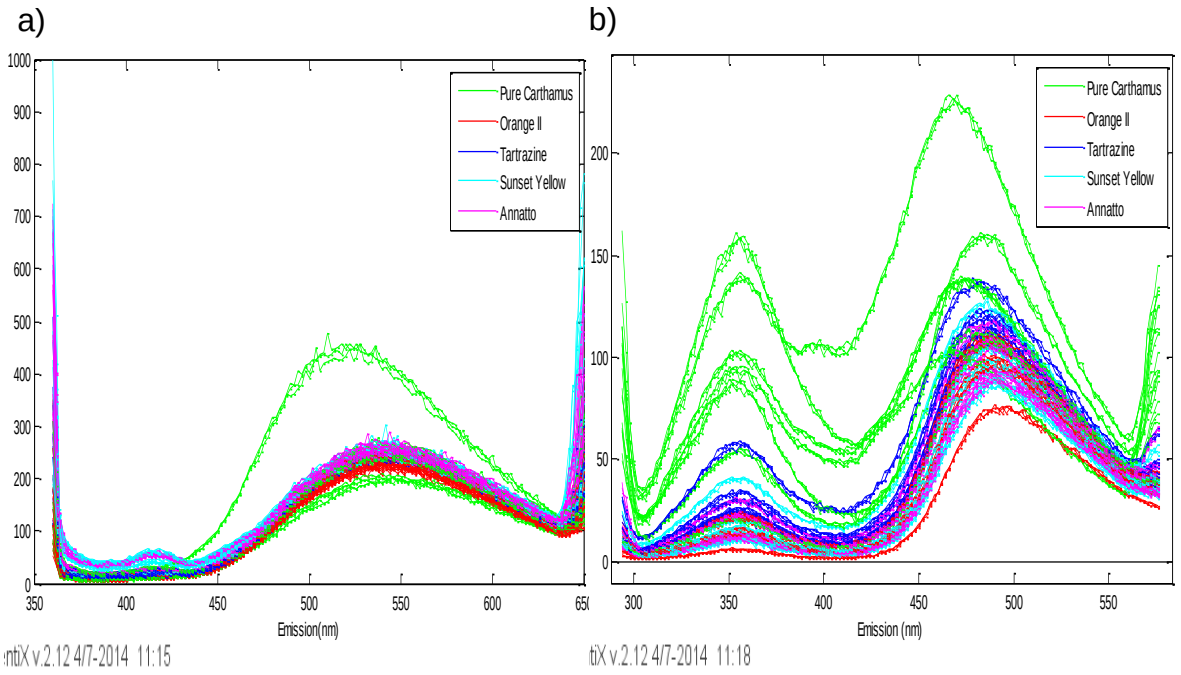


Figure 4.11 Raw spectra for the screening phase using Fluorescence spectroscopy. A) in front-face mode (Excitation region: 340 nm); b) in right-angle mode (Excitation region: 290 nm).

NEAR-INFRARED

Figure 4.12 shows the NIR raw spectra for the screening phase containing all the pure *Carthamus* samples and all the samples tampered with unwanted additives. By briefly analyzing this plot it can be seen that all the samples show a quite similar behavior and, in that sense, it would not be easy to use this data to create a model capable of distinguish the pure from the impure samples. These results indicate that the NIR data might be irrelevant for future multivariate analysis.

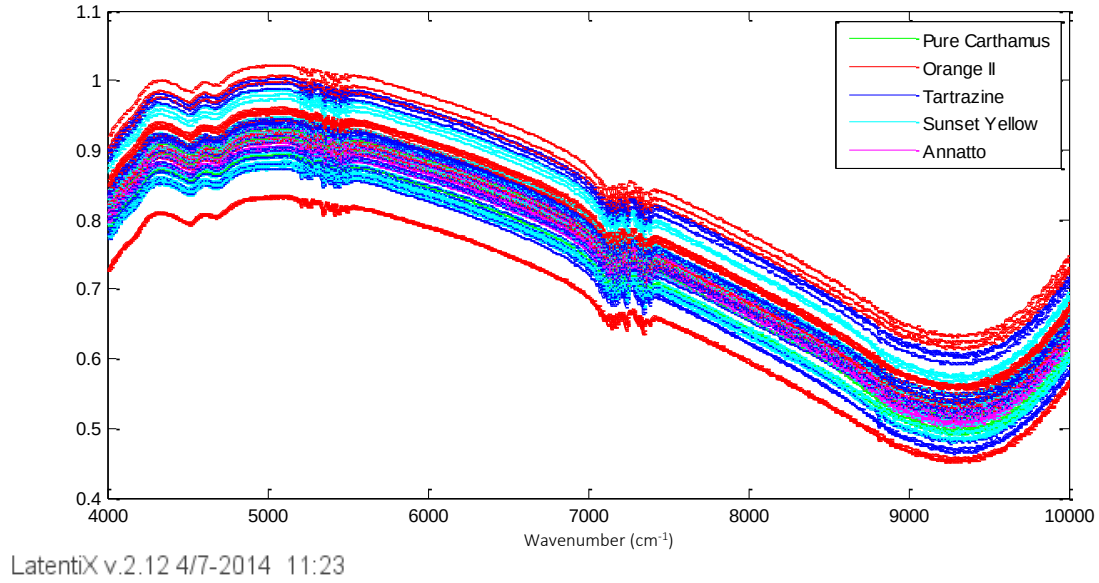


Figure 4.12 Raw spectra for the screening phase using NIR. (Resolution: 8cm^{-1}; 64 scans)

4.4. Multivariate Data Analysis

4.4.1. EXPLORATORY ANALYSIS

This section has the purpose of exploring and understanding the data by projecting it in visual graphs, which enables a much easier interpretation of the results. In this case, the projection technique used was Principal Component Analysis (PCA) since it is simple and is the most commonly used technique.

The PCA was performed using LatentiX and the results obtained and further discussion for each spectroscopic method are shown below. A brief description of the process for detecting outliers in the data is also presented.

UV-VISIBLE

The first step on the analysis of UV-Vis data was to perform a PCA on all the UV-Visible data. Subsequently, the presence of possible outliers was studied by analyzing the scores plot (Figure 4.13 a)) and the Influence plot – Hotelling's T^2 vs Q-residual (Figure 4.13 b)). The two samples that have been consider as outliers were the samples containing 1.6 mg of Orange II and 0.8 mg of Tartrazine per kg of pure *Carthamus* (each with 3 replicates, giving 6 objects - marked in Figure 4.13). Tartrazine sample is not identified as an extreme outlier in the influence plot, but by further analyzing the scores plot it was concluded that it should be defined as an outlier.

The PCA scores plot containing only the data considered as reliable for the model (without outliers) is presented in Figure 4.14 a). Additionally, Figure 4.14 b) shows the scores plot for the PCA model containing the data with all the pure *Carthamus* samples, since the main objective of this work is to create a model that is able to distinguish the adulterated samples from all types of *Carthamus*.

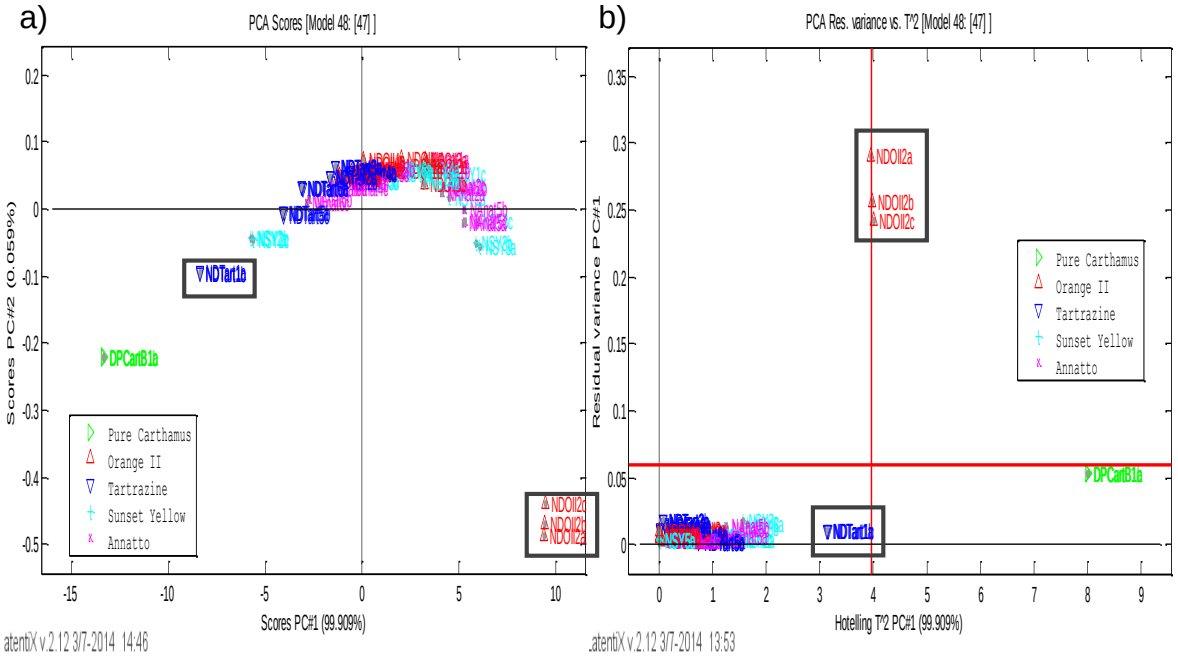


Figure 4.13 a) Scores plot for the UV-Visible spectroscopy data; b) Influence plot (Hotelling's T^2 vs Q-Residual) for the UV-Visible spectroscopy data with the critical limits shown in red; The dark grey squares show the samples defined as outliers (Data Pre-processing: Mean Centering).

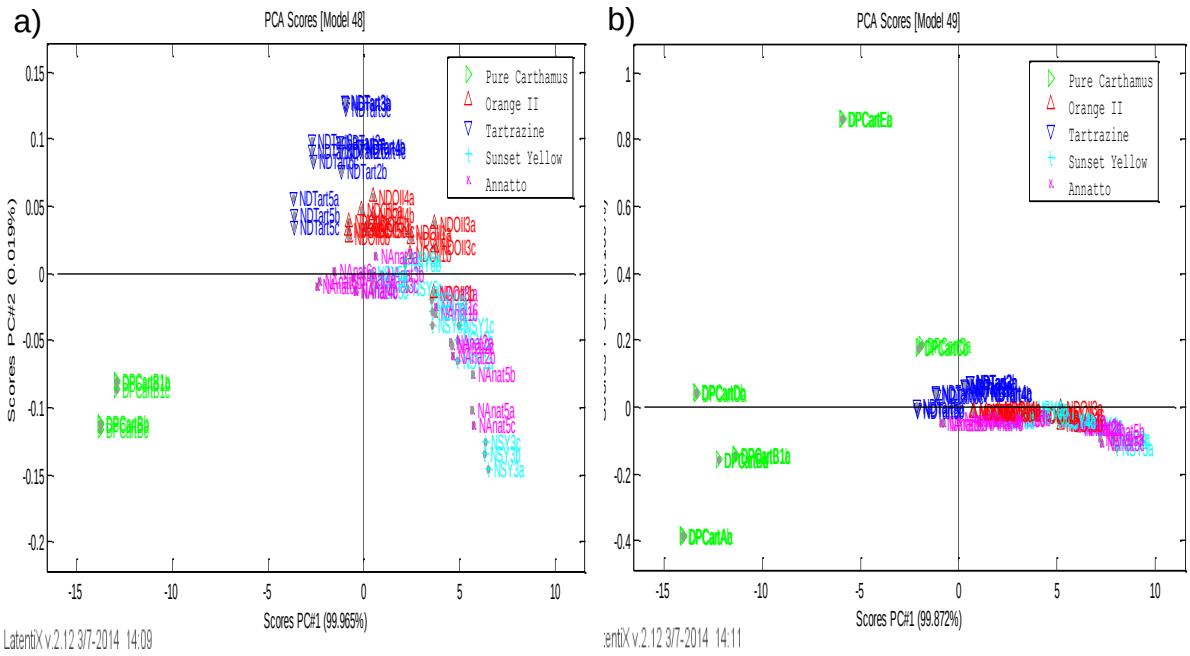


Figure 4.14 PCA scores plots for the data from UV-Vis spectrophotometer after removing the outliers) a) containing only the information for pure Carthamus B and all the adulterated samples; b) containing all the samples for pure Carthamus and all the adulterated samples; (Pre-processing: Mean Centering).

Even though the identification of each unwanted additives present in *Carthamus* does not seem possible, it can be seen a clear separation between the different impure samples and the pure *Carthamus,* which is the main purpose of this project.

The Loadings plot (Figure 4.15) for the first principal component indicates that the region 320 to 420 nm is the one that gives more information about the data, as it was already observed in the raw spectra. For this reason it was also performed the PCA using only these region but it did not showed superior results. The results obtained are presented in Appendix III.

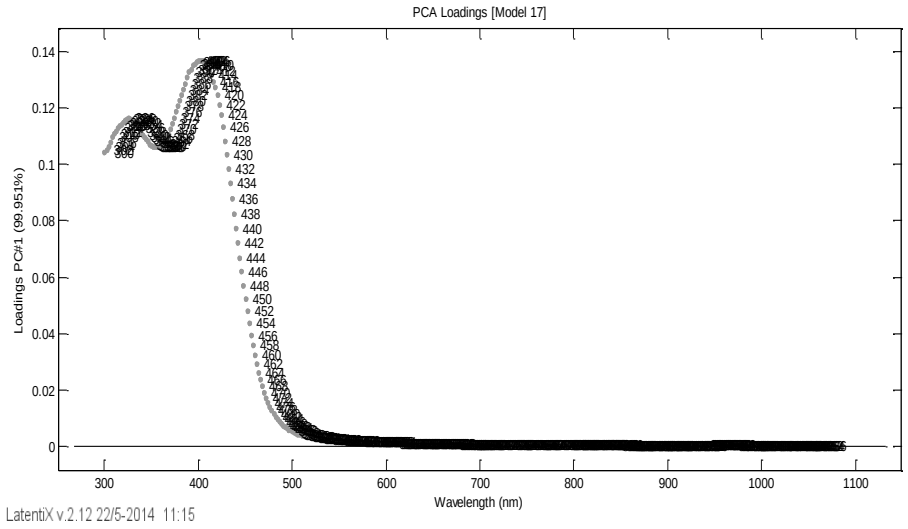


Figure 4.15 PCA loadings plot for the data from UV-Vis spectrophotometer (Pre-processing of raw data: Mean Centering).

FLUORESCENCE

An exploratory analysis was also performed in the fluorescence data, using PCA, in order to determine what instrument which would be the best to use in further analysis. PCA was implemented in the data obtained in front-face and right-angle mode.

Firstly, for the data obtained in front-face mode, a PCA with all the data measured in an EEM was performed. By examining the scores plot (Figure 4.16 a)) and the influence plot (Figure 4.16 b)), it is possible to detect outliers which were the 3 replicates for the *Carthamus* containing 8 mg of Sunset Yellow per kg of pure *Carthamus* and one of the replicates for the *Carthamus* sample containing 0.8 mg of Sunset Yellow per kg of pure *Carthamus* (both identified in Figure 4.16).

Figure 4.17 a) and b) show the results for PCA after excluding the outliers, the first plot is the scores plot for the data containing only the *Carthamus B* pure and the impure samples and the second plot is the scores plot obtained for the PCA model made including also the remaining pure *Carthamus* samples. Considering these two plots it can be observed a slight separation between the pure *Carthamus* and the impure ones and, therefore, a further analysis could be done in order to verify that the information given by this data is useful. However, at a first sight, it does not seem to present better results than the UV-Visible data.

Considering the previously presented spectra for the screening phase (Figure 4.11 a)) that showed a region where a distinction between pure and impure samples seemed likely, a PCA was also performed for this specific region (370 to 450 nm). The resulting PCA scores plot is displayed in Figure

4.18. By analyzing this figure, it cannot be observed any superior results comparing to the previously obtained.

Even though the spectra acquired through Fluorescence spectroscopy do not appear to have very promising results, it is known that the setup and settings used in the instrument influence the resolution of the spectra, so it could be interesting to further explore different setups and setting in this instrument.

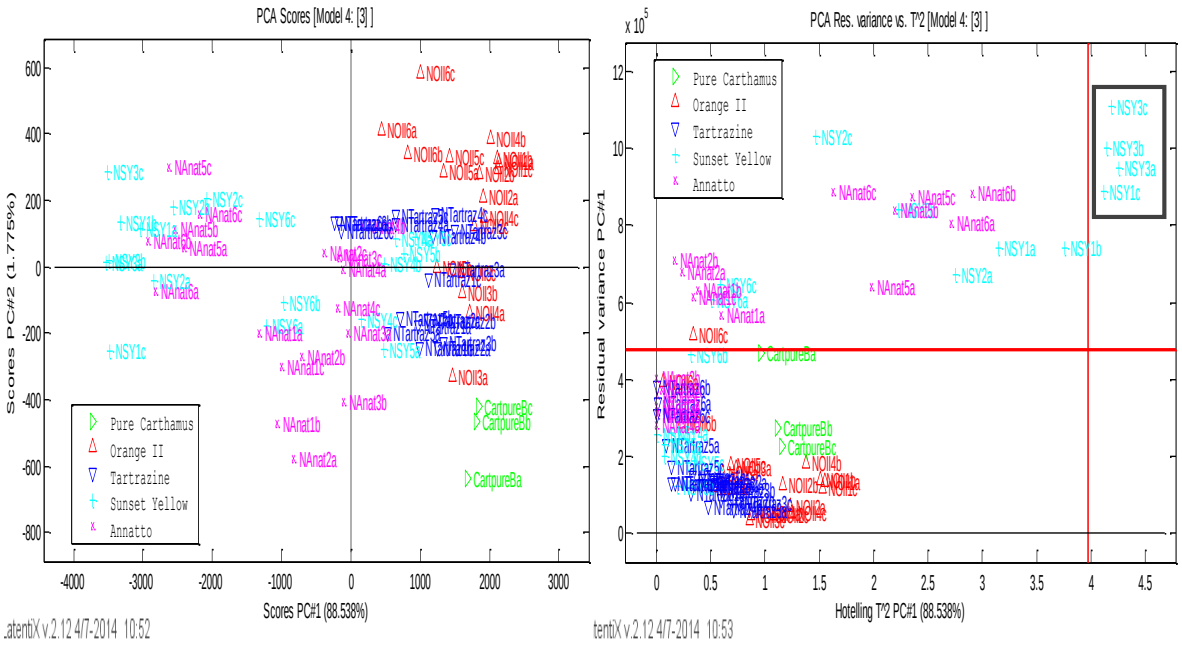


Figure 4.16 a) Scores plot for the Fluorescence spectroscopy data, measured in front-face mode; b) Influence plot (Hotelling's T^2 vs Q-Residual) for the Fluorescence spectroscopy data, measured in front-face mode, with the critical limits shown in red. The dark grey squares show the samples defined as outliers; (Data pre-processing: Mean Centering).

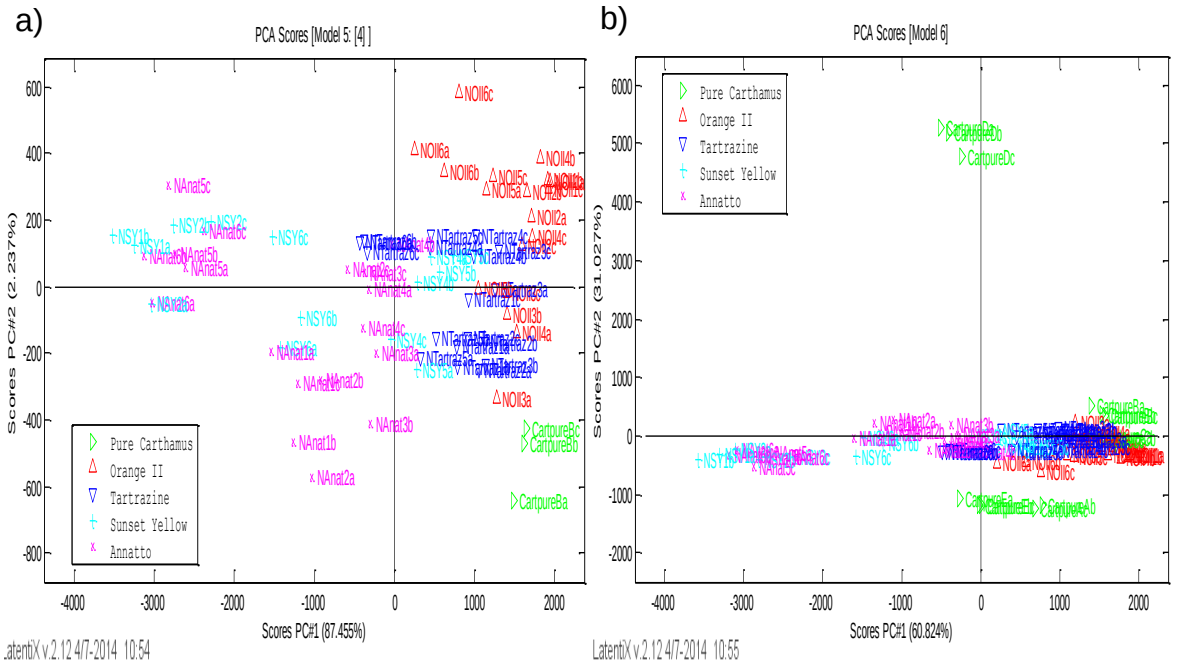


Figure 4.17 PCA scores plots for the data from Fluorescence spectrophotometer measured in front-face mode, after removing the outliers (Pre-processing: Mean Centering); a) containing only the information for pure Carthamus B and all the adulterated samples; b) Containing all the samples for pure Carthamus and all the adulterated samples.

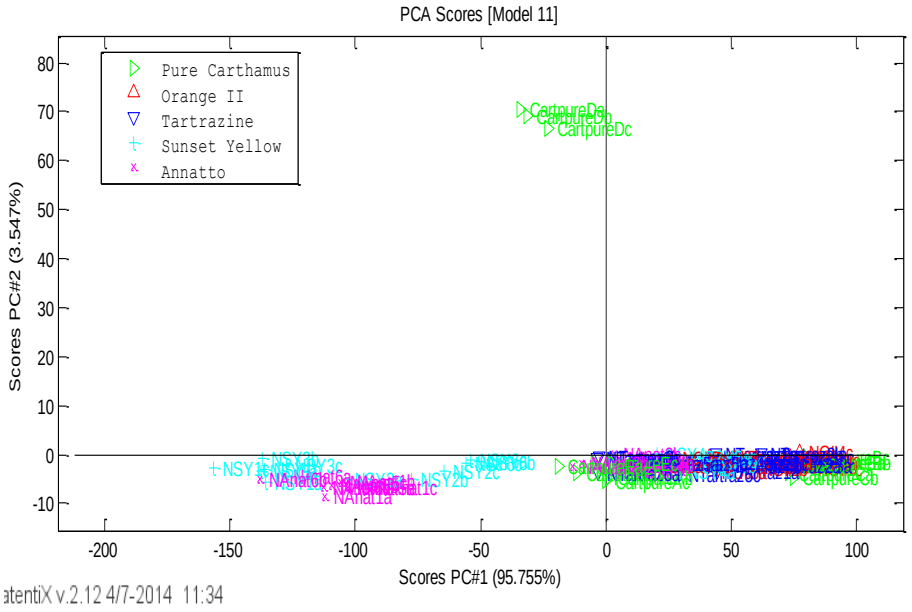


Figure 4.18 PCA scores plots for the data from Fluorescence spectrophotometer measured in frontface mode for the emission region from 370 to 450 nm (Excitation region: 340 nm).

Likewise, a PCA was also performed using all the data obtained in right-angle mode. The resulting scores plot is presented in Figure 4.19 a).

Through an analysis of this scores plot (Figure 4.19 a)) and the influence plot (Figure 4.19 b)) it was possible to detect the *Carthamus* sample containing 0.8 mg of Tartrazine per kg of pure *Carthamus* as an outlier. This was the exact same sample used in UV-Visible, where this sample was also consider an outlier.

Hence, Figure 4.20 a) and b) show the results for the PCA using the data considered as trustworthy, after removing the outliers. While the plot on the left includes only the pure *Carthamus B* and the impure samples, the plot on the right was obtained with the PCA model that includes also the remaining pure *Carthamus* samples.

By looking at these two plots it can be observed that there is no clear separation between each of the unwanted additives and between the pure and impure samples which demonstrates that this data could not be proper to analyze in future studies. However, and as it has already been mentioned, it would be interesting to understand how the setup and settings used in the fluorescence spectrophotometer could influence the results previously presented.

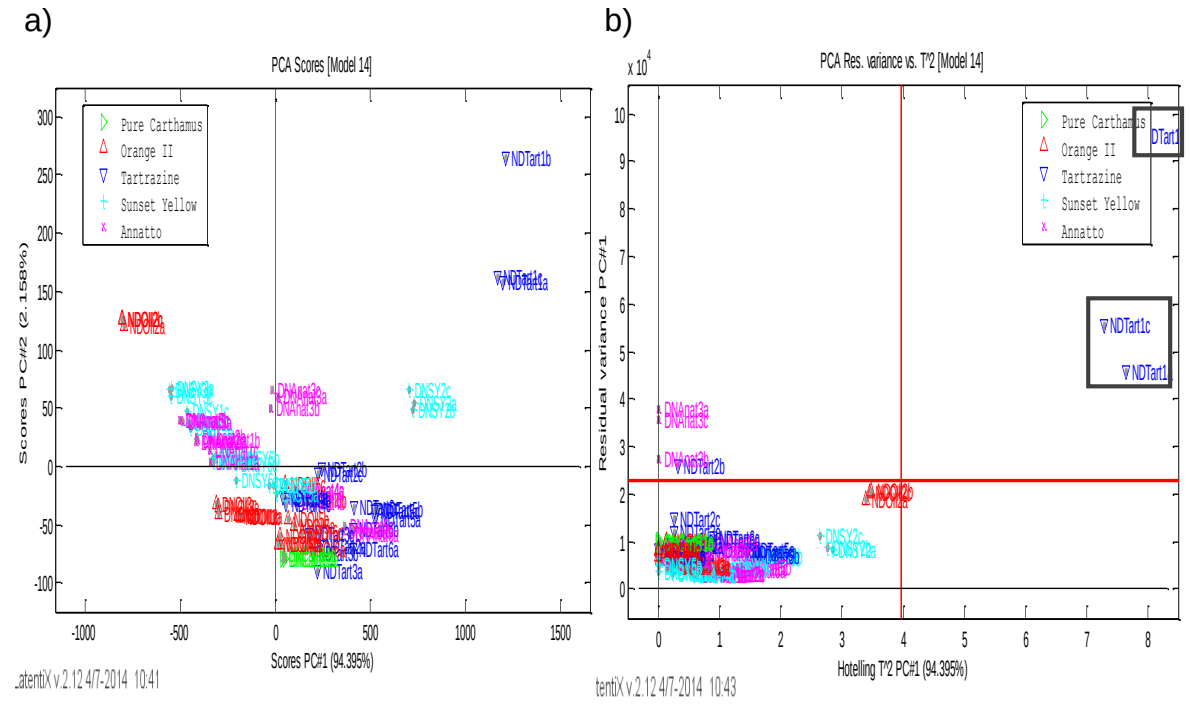


Figure 4.19 a) Scores plot for the Fluorescence spectroscopy data, measured in right-angle mode; b) Influence plot (Hotelling's T^2 vs Q-Residual) for the Fluorescence spectroscopy data, measured in right-angle mode, with the critical limits shown in red. The dark grey squares show the samples defined as outliers; (Data pre-processing: Mean Centering).

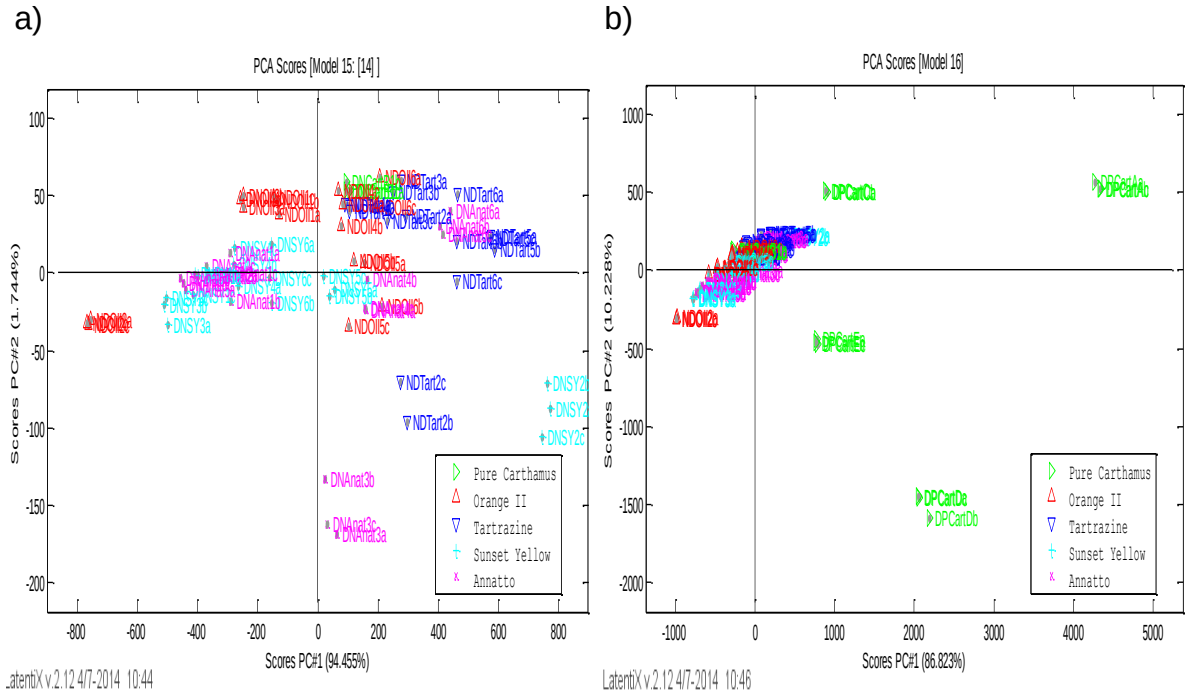


Figure 4.20 PCA scores plots for the data from Fluorescence spectrophotometer measured in right-angle mode, after removing the outliers (Pre-processing: Mean Centering); a) containing only the information for pure Carthamus B and all the adulterated samples; b) containing all the samples for pure Carthamus and all the adulterated samples.

NEAR-INFRARED

A PCA was also performed in the NIR data and the results obtained are shown in Figure 4.21. These results are not relevant since they do not show any separation or tendency worth analyzing. The fact that the limit of detection for NIR spectroscopy is usually about 0.1% [55] might be the reason for the results obtained with this method, since it does not have the ability to detect compounds in such low quantities as the ones handled in this research. Therefore, the data obtained in NIR is not useful in this specific situation since it gives significantly worse results than the other methods used.

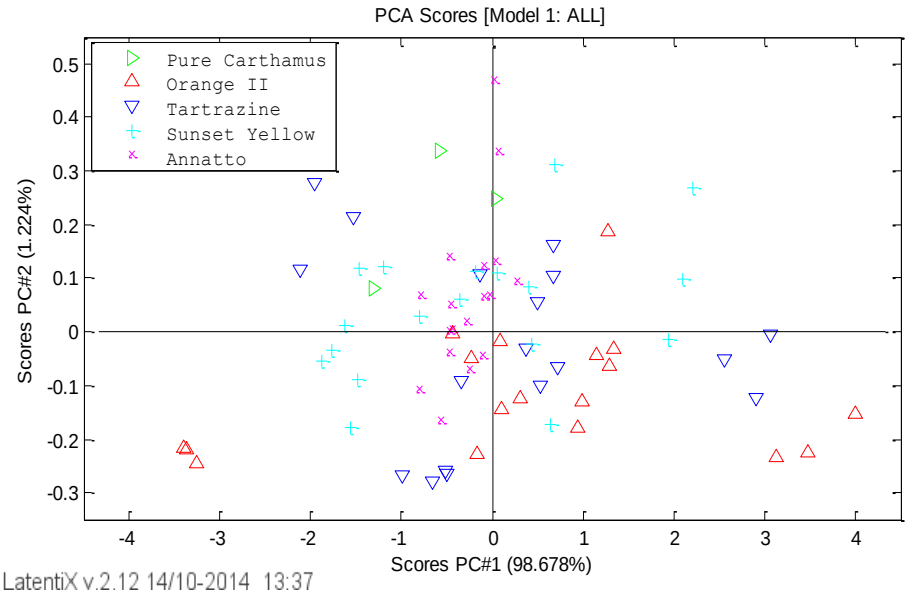


Figure 4.21 PCA score plot for the data from NIR spectrophotometer (Pre-processing: Mean Centering).

4.4.2. UV-VISIBLE DATA

QUANTITATIVE ANALYSIS

By analyzing the raw spectra and the results acquired with the exploratory analysis it became apparent that UV-Visible data is the one that shows the most interesting and useful results. Hence, this data will be the first to be further analyzed.

With the data obtained in the screening trial it was expected to create a model that could quantify the unwanted coloring agents added to the natural color through Chemometric tools using LatentiX *V 2.12*. In order to do that, a Partial Least Square (PLS) model was created for each of the unwanted color additives in which the Y variable, i.e. the predictive response, is the concentration of unwanted additive added to the natural coloring product *Carthamus*.

Therefore, with the UV-Vis data previously preprocessed using mean centering, four PLS models were created in LatentiX. These models were cross-validated by using a variable containing the information about the replicates for each sample, placing these replicates in the same cross-validation segment. The cross-validation method used was the leave-one-out method (full CV).

The PLS scores plots for all four models are displayed in Figure 4.22 and Figure 4.23 and the content of these figures are colored according to the concentration of unwanted additive, together with the corresponding plots for the RMSE according to the number of latent variables (LV).

By looking at these figures, it seems to be difficult to quantify the unwanted additives since the different concentrations levels do not show a linear distribution in the score plot obtained.

Besides that, by looking at the plot for the RMSE according to the number of LVs it can be observed that for Orange II, Tartrazine and Sunset Yellow (the synthetic compounds) there is a good trend, however the RMSE for all the cases is very high (> 30 ppm). For Annatto, there is an odd trend for the RMSE which shows that this compound might have some problems to be detected in the natural coloring product because of its high dilution and also the fact that this is a natural product, as seen before.

Therefore, the creation of a good and trustworthy PLS model for quantification using this data is not possible. It is important to point out that other PC scores were also analyzed but PC1 vs PC2 showed the best tendency.

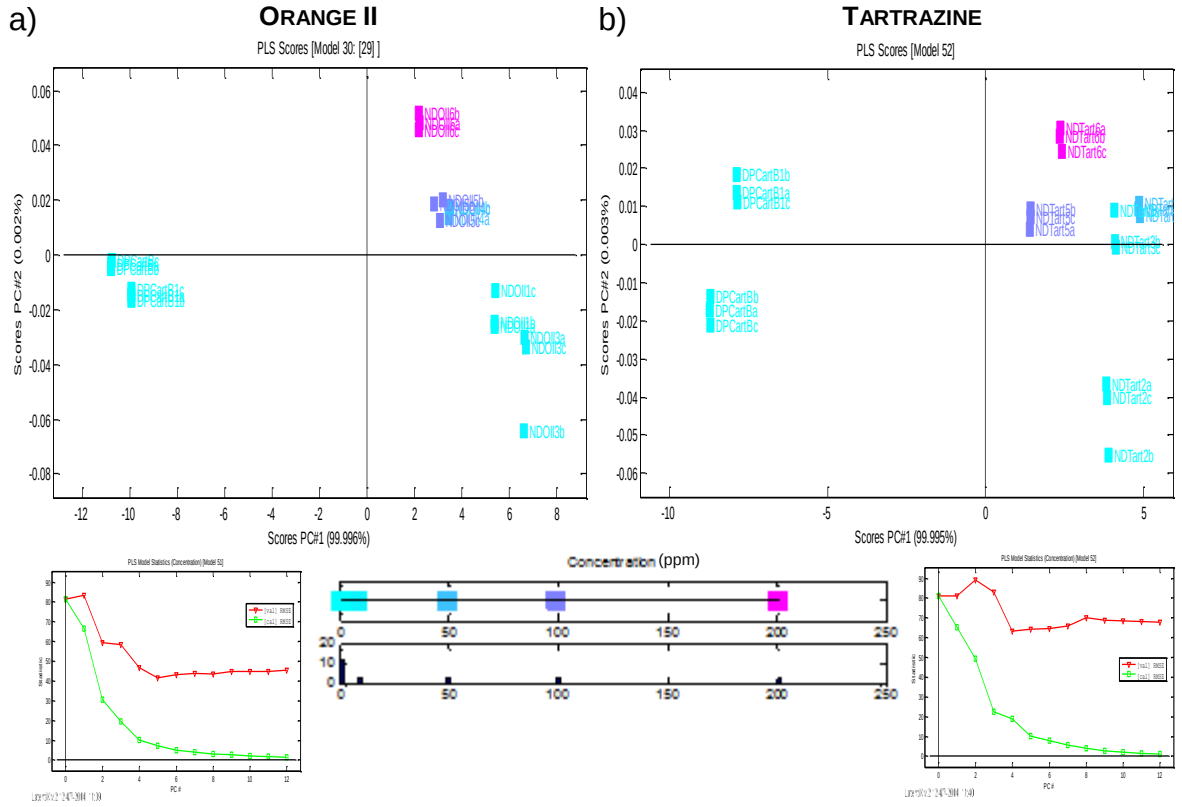


Figure 4.22 PLS Scores plot colored according to the concentration (in ppm) of the unwanted additive and the corresponding RMSE plot showing the RMSEC (green line) and RMSECV (red line) according to the number of LVs; a) for the unwanted additive Orange II; b) for the unwanted additive Tartrazine.

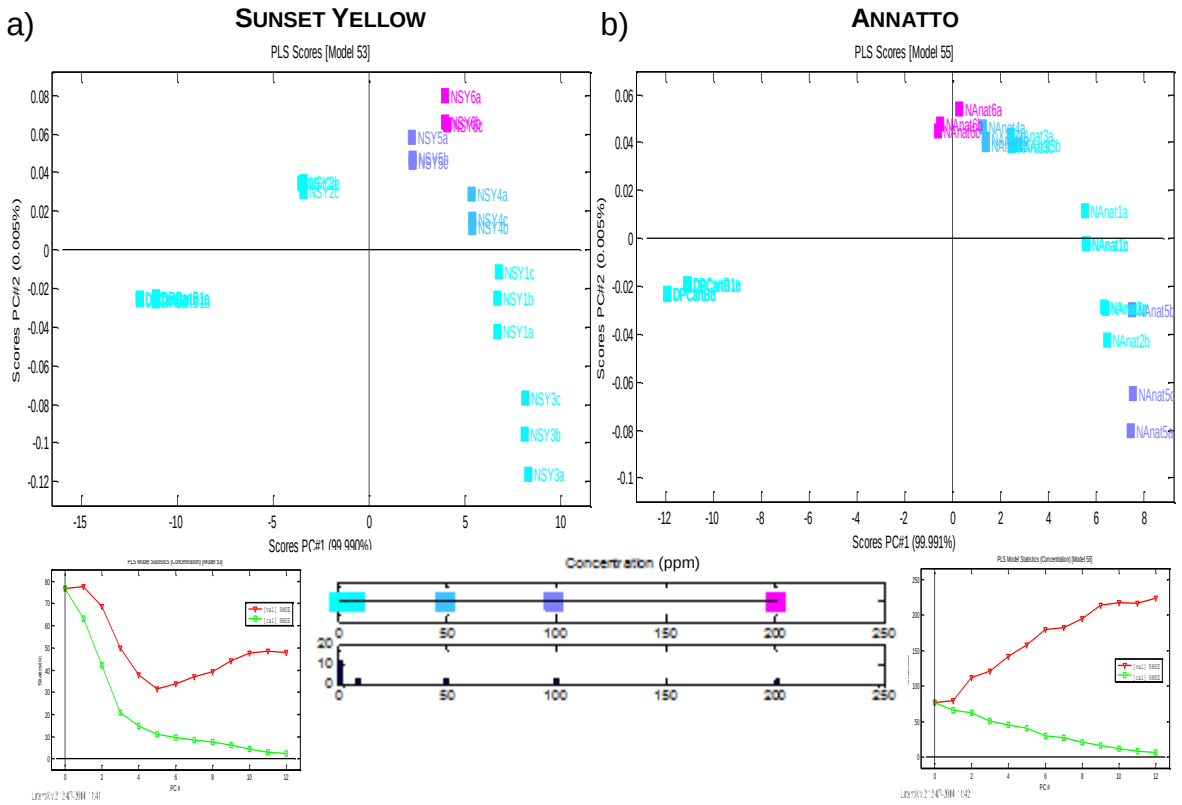


Figure 4.23 PLS Scores plot colored according to the concentration (in ppm) of the unwanted additive and the corresponding RMSE plot showing the RMSEC (green line) and RMSECV (red line) according to the number of LVs; a) for the unwanted additive Sunset Yellow; b) for the unwanted additive Annatto.

QUALITATIVE ANALYSIS

As was mentioned previously, the data obtained in the screening trial phase for UV-Visible seems to be the one that can be used to create a model that could distinguish the pure *Carthamus* samples from the impure ones. In order to do that, a classification method should be created.

The classification methods used, that seemed appropriate in this particular case, were Partial Least Squares Discriminant Analysis (PLS-DA) and Soft Independent Modeling of Class Analogies (SIMCA). With the aim of obtaining better and trustworthy results, more *Carthamus* samples were created. Therefore, all the five *Carthamus* were used in the data and, in addition, a set of mixtures of the five *Carthamus* was also used. From the set of 22 mixtures made with the pure *Carthamus* products, only 7 were used since the other ones showed a lot of noise caused by the saturation of the instrument detector.

To perform the PLS-DA method, a classification model is created through regression of the matrix containing the spectral data against one matrix containing the information for the classes (through a binary code), using the PLS algorithm. This model is validated by cross-validation using the variable that refers to the replicates, the same way as for the PLS models for the quantitative analysis. It is important to stress that the only preprocessing method used was Mean Centering since no scattering effect were detected. However, other preprocessing methods were tested, e.g. Auto Scaling and Standard Normal Variate (SNV), but they did not show superior results.

The PLS-DA method was applied with the purpose of distinguish two categories - Pure *Carthamus* and Impure and the results obtained are presented and discussed below.

In Figure 4.24 a), the PLS scores plot for the model created using the UV-Visible data is presented. The data present in this plot contains the five pure *Carthamus* supplied by Chr. Hansen and the impure samples. By plotting the scores for the first against the second principal component, a clear separation between the pure and the impure samples is shown. Figure 4.24 b) illustrates the PLS scores plot for the data with the additional set of combinations of *Carthamus*. This plot also shows a good separation between the two categories and, thereby, this will be the model used in the classification task since it has more information about the Pure *Carthamus* class. However, it should be stressed that in this second case (Figure 4.24 b)), PLS-DA, as a linear classification method, is not completely efficient since the separation of the two classes is not linear.

The plot showing the RMSE for calibration and validation of the model according to the number of LVs is presented in Figure 4.25. By analyzing this plot the optimal number of latent variables was selected, which was 6 LVs, since the corresponding RMSECV is the minimum of the curve.

In order to analyze how PLS-DA classification is working and how successful was the method in classifying the samples, a confusion matrix can be arranged, in which the numbers of well classified and misclassified samples for each category are presented in a 2x2 matrix. The confusion matrix for the present case is presented in Table 4.1.

By looking at the confusion matrix it can be observed that only 3 of the pure samples were misclassified by the model as impure which indicates that the model is quite efficient in the classification process. It is relevant to enhance the fact that the absence of impure samples that have been misclassified as pure (false negatives for the impure class), makes this model more valuable and trustworthy, especially for its implementation in the food industry.

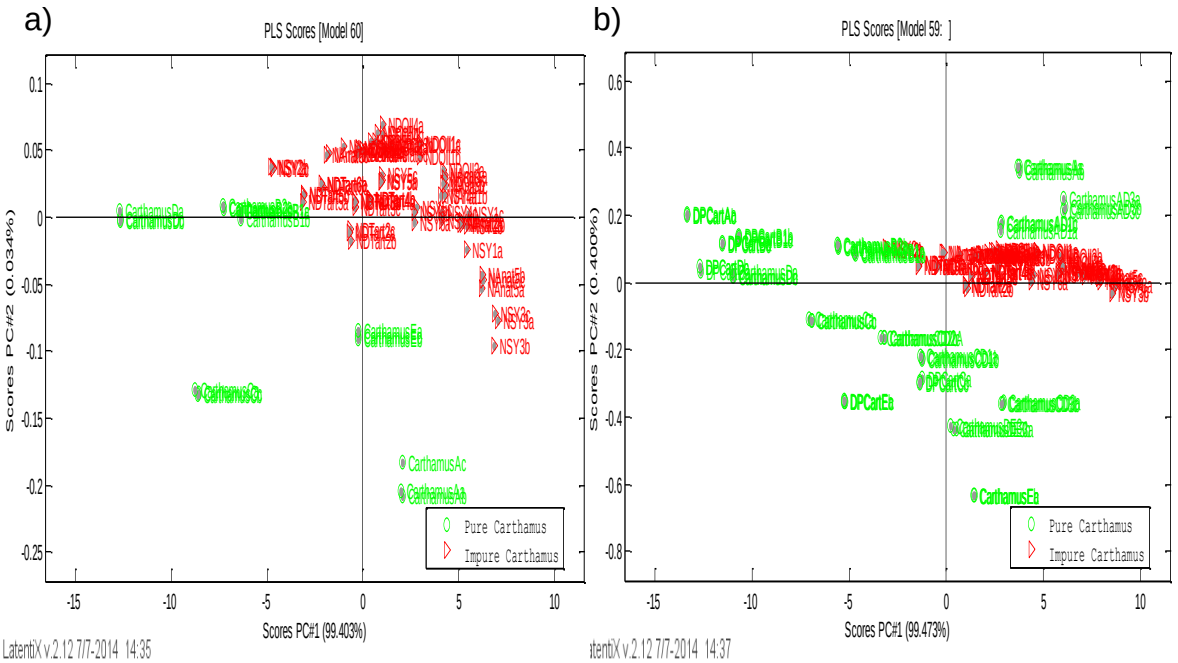


Figure 4.24 Scores plot for the data from UV-Vis colored according to the two categories: Pure and Impure Carthamus (Pre-processing of raw data: Mean Centering; Dilution: 1000 ppm); a) including the Carthamus pure and impure samples; b) including the Carthamus pure and the additional set of combinations of the five pure Carthamus and the impure samples.

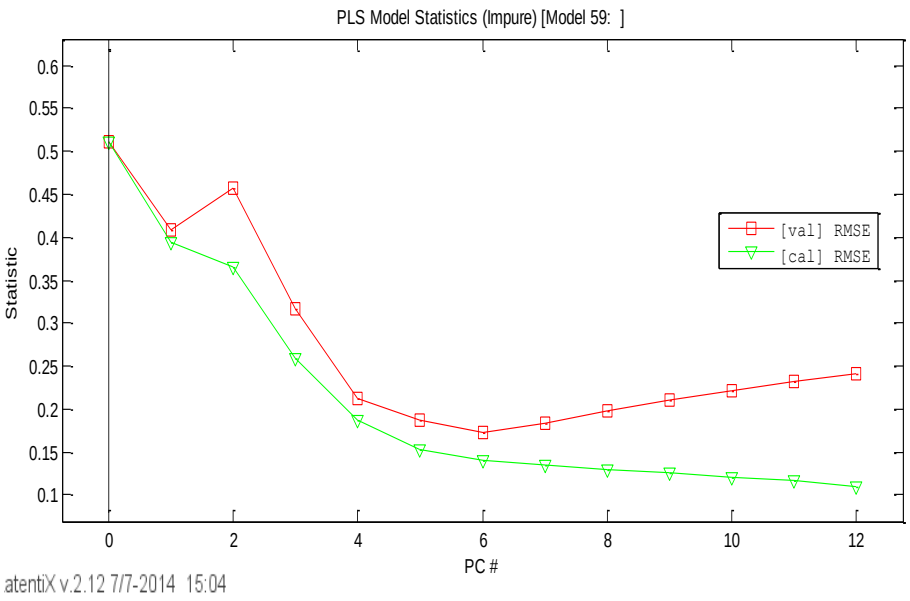


Figure 4.25 RMSE for Calibration (green inverted triangles) and for Cross-Validation (red squares) intended to the Impure Carthamus class against the number of latent variables.

Table 4.1 PLS-DA classification confusion matrix for the UV-Visible data.

		Pure	**Impure**
Predicted as	Pure	**54**	0
	Impure	3	**66**

Besides PLS-DA, SIMCA was also implemented, in order to study if better results could be achieved. This classification model is performed through the creation of a PCA model for each class, in this case, two PCA models were created for the two classes, pure and impure *Carthamus*. Since there is only two categories there is only the need to use one of the classes. Therefore, it was used the data for the pure *Carthamus* samples and, subsequently, the impure *Carthamus* samples were predicted. It should be noted that cross-validation has been used in the PCA model.

After analyzing the loadings plot and the model statistics it was concluded that the optimal number of components for this model was three. Therefore, the resulting plot showing the *Q*-Residual against the Hotelling's T^2 (Influence plot), for three components, is presented in Figure 4.26.

Through an analysis of this plot, it can be observed that the impure samples (red markers), are separated from the pure samples (green markers), but still the samples contaminated with Tartrazine are quite close to the pure samples. Moreover, the model validation was not completely satisfactory, as it can be seen by looking at the validated pure samples (blue markers) in Figure 4.26. Additionally, the confusion matrix is also presented in Table 4.2, in which it is possible to verify that this model misclassifies 15 pure samples and 3 impure samples.

Unfortunately, this result using the SIMCA method is not very satisfactory when compared with the model obtained with PLS-DA, since the latter shows a superior result in terms of classification performance.

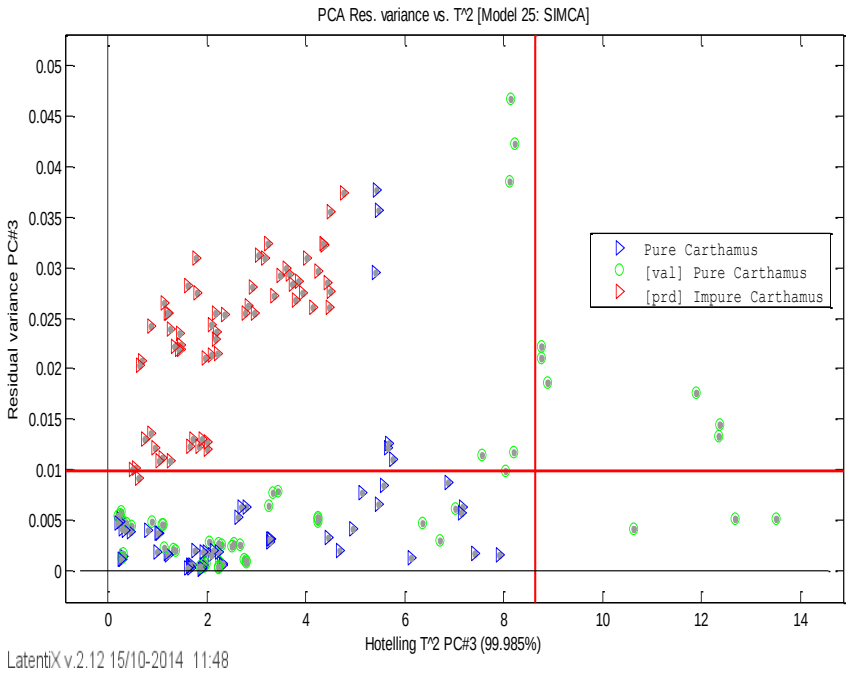


Figure 4.26 Influence plot (Hotelling's T^2 vs Q-Residual) for the PCA model regarding SIMCA classification (The labels for each point were removed to facilitate the visualization)

Table 4.2 SIMCA classification confusion matrix for the UV-Visible data.

		Pure	**Impure**
Predicted as	Pure	**42**	3
	Impure	15	**63**

Since the results obtained so far did not completely comply with the objectives of the present project, other strategies were studied, such as changing the dilution ratio when diluting the sample in water and using a more restricted spectral range.

As mentioned above, the dilution ratio used in the samples for the screening phase was 1000 ppm in water, however it was observed that in some parts of the spectra obtained there was some noise. This led to the assumption that there might be problems caused by the saturation of the detector from the UV-Visible spectrophotometer and, therefore, a further dilution (10 ppb of mixture in water) was made to understand if there was more information being omitted by these detection issues. The same data analysis was made for this data, but, unfortunately, the results obtained were not satisfactory. These results are shown in Appendix IV.

Additionally, the spectral range used in the construction of the classification model was studied and, hence, a model using a specific range was created to comprehend if it could offer a better model than for the full range previously used. The specific spectral region studied was from 320 to 420 nm, the one previously mentioned in section 4.2, as being the region in which the pure *Carthamus* samples and the pure unwanted additives showed different behaviors. Once again, the resulting model was inferior to the first model obtained using the full spectral range.

It also important to note that it was also performed PLS-DA and SIMCA to classify the samples for the 5 different classes – Pure *Carthamus* and the four classes corresponding to *Carthamus*

containing the four different unwanted additives -, however, as expected the results obtained were not satisfactory. The results for the PLS-DA performed for the 5 classes using the two different sample dilutions are presented in Appendix V.

Even though the results for the classification between pure and impure samples are quite good, especially when using PLS-DA, the remaining results, regarding the development of models to identify and quantify the unwanted color additives in *Carthamus*, did not show satisfactory results.

For that reason, a linearity test was made to check if the concentration of unwanted additives in each sample was linear or else some mistake has been made in the previous measurements or the instrument was not working properly.

Additionally, it is assumed that one of the reasons for these unfavorable results is the fact that the dilution of the unwanted additives in 1:1000 proportion in water, necessary to properly measure the samples in UV-Vis spectrophotometer, results in a very low quantity of color additives added to the natural color and, consequently, in a more difficult detection of these compounds. Therefore, it would be interesting to see if the same quantity of additive in water could be detected by the instrument.

In order to do this test, the three synthetic additives – Orange II, Tartrazine and Sunset Yellow -,were added to water in 7 different concentration levels (160 ppb, 4 ppm, 8, ppm, 16 ppm, 40 ppm, 80 ppm and 160 ppm). The concentration levels in ppm were made to simulate a similar situation as the one previously made using *Carthamus* as solvent in the screening phase. This experimental trial allows to verify that there is linearity in the spectra obtained and, additionally, to understand if the same quantity of illegal additives could be detected in water. To the latter case, a ppb level was also used, simulating the same quantity of unwanted additive in *Carthamus* for the 160 mg pure additive / kg *Carthamus* concentration level after the dilution in water in a 1: 1000 ratio.

The test was only made for the illegal additives, since Annatto showed such a different behavior in the pure spectra caused by the fact that this is a natural product and it has been previously diluted, thus, the resulting spectra would be of lower intensity of the remaining additives and, therefore, it is not necessary to use Annatto in the test.

The resulting spectra for the linearity test are presented in Figure 4.27 and Figure 4.28.

After observing Figure 4.27, it is completely evident the linear behavior for what concerns the concentration of unwanted color additives for all three additives, which verifies that the sample preparation was made correctly and the instrument is working properly.

On the other hand, it can be seen in Figure 4.27 and, in detail, in Figure 4.28 that for the concentration of 160 ppb the signal is quite low, almost imperceptible.

This indicates that, in a similar way, the quantities of unwanted additives added into the natural food color *Carthamus* in the screening phase are too low to be significantly detected in the UV-Visible spectrophotometer. However, according to the results obtained for the classification model using PLS-DA, the separation between *Carthamus* pure and impure is successfully achieved, which indicates that the minor influence of the additives has an effect in the spectra for the pure *Carthamus*.

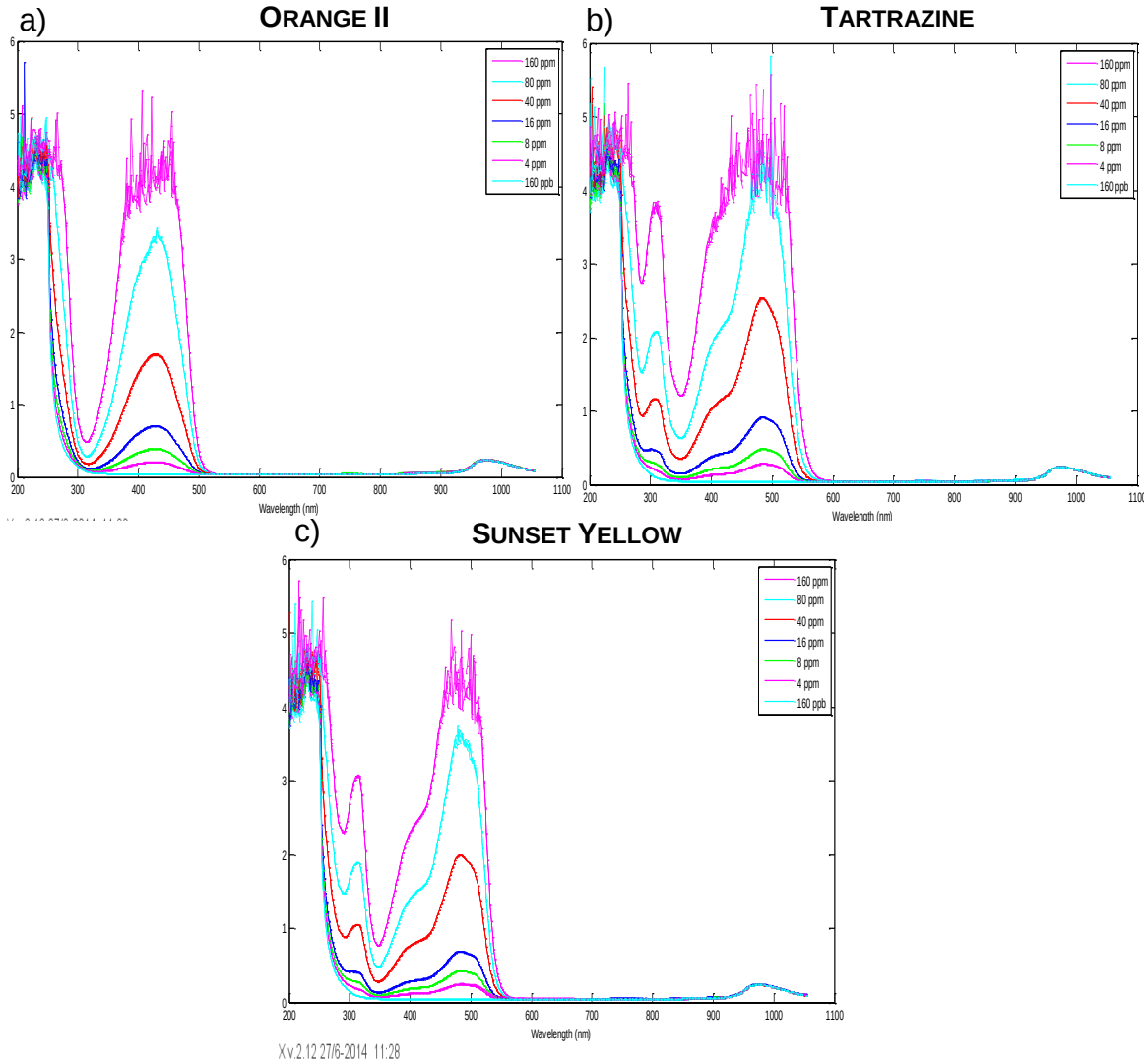


Figure 4.27 Raw spectra for the linearity and detection tests made using 7 different concentrations levels of the 3 synthetic additives a) Orange II; b) Tartrazine; c) Sunset Yellow.

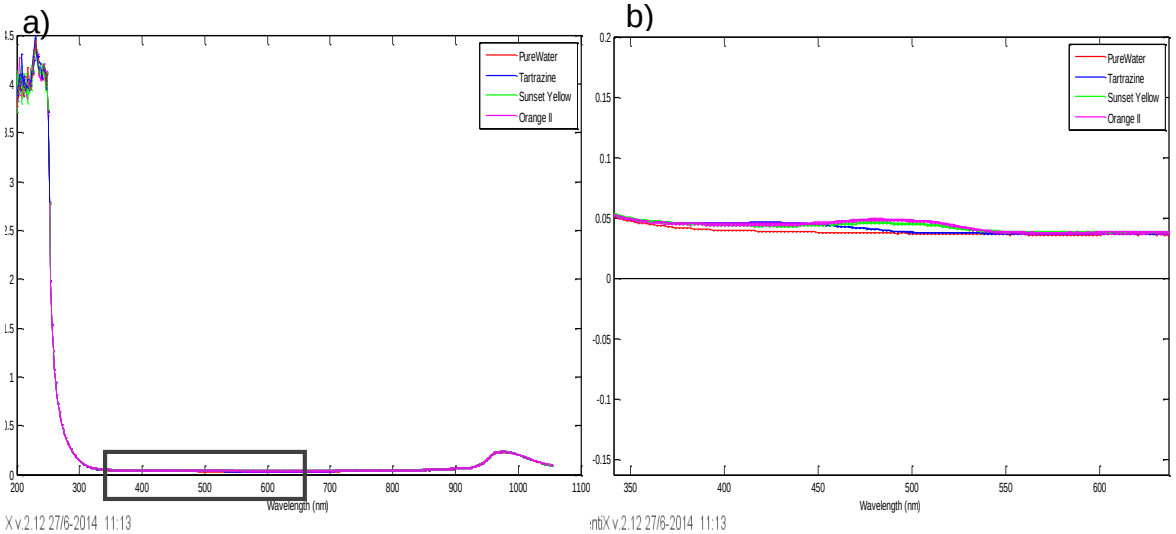


Figure 4.28 Raw spectra for the detection test made for the four water samples: one pure and three contaminated with 160 ppb of the 3 synthetic additives – Orange II, Tartrazine and Sunset Yellow; b) Close up of zone marked with a square in the figure on the left.

4.4.3. FLUORESCENCE DATA

As has been mentioned earlier, the outcome from the results for the UV-Visible data was not completely satisfactory and, for that reason, a further evaluation of the data from fluorescence spectroscopic was made.

Through a deeper analysis of the fluorescence data previously obtained, it was observed that by using that specific setup and settings in this instrument in front-face mode there is a high probability of occurring a phenomenon known as Fluorescence Resonance Energy Transfer (FRET), which might be influencing the resolution of the spectra obtained. In this particular case, what might be happening is that the fluorescence signal of the unwanted additives is being reabsorbed by the *Carthamus* signal and, thus, it cannot be detected.

In order to solve this issue, the light path length in the instrument can be decreased, by moving the sample holder, which could drastically reduce the occurrence of FRET and, therefore, improve the resolution in the spectra acquired. In addition, the dilution of the samples in water, rather than measuring them in pure form - as it was previously done for front-face mode -, together with the optimization of the instrument settings can also help to increase the signal resolution.

Accordingly, different trials were made, in which three factors were being studied: the dilution proportion in water used in the samples, the instrument settings and the sample holder position. The objective was to combine the most appropriated situation for the three factors involved. First, a chosen sample was diluted in water in 4 different dilution ratios (1:10 000; 1:1000; 1:100; 1:10), then different settings (Spectral range, scan control, scan rate and detector voltage) were experimented for the four samples. Additionally, the sample holder was placed closer to the light source in an appropriate distance, after testing different positions.

After this evaluation, the conclusion is that the best dilution proportion is 1:1000 and the appropriated settings are the ones presented in section 3.5.

EXPLORATORY ANALYSIS

Once all the data from the fluorescence instrument was acquired, the next step was to analyze through PCA. Firstly, the presence of outliers was investigated, by analyzing the Hotelling's T^2 vs Q-residual plot and the PCA scores plot. Thus, two samples were identified as outliers: one of the replicates for the *Carthamus* containing 8 mg of Tartrazine per kg of pure *Carthamus* and one of the replicates of *Carthamus* containing 0.8 mg of Annato per kg of pure *Carthamus*. After the removal of the outliers, PCA was performed and the corresponding scores plot is presented in Figure 4.29.

In Figure 4.29 a), it is possible to identify a reasonable separation between the pure *Carthamus* and the samples containing the different unwanted additives, but still the samples containing Tartrazine and Sunset Yellow are overlapping each other. For that reason, the focus will be in analyzing the separation between the pure and impure samples of *Carthamus*, being this the main goal of this project.

Therefore, by looking at Figure 4.29 b), containing not only the impure *Carthamus* samples, but also all the five pure *Carthamus* samples, a separation between pure and impure samples is easily recognized, even though this appears to be a non-linear separation. Nonetheless, through a deeper analysis of this data this can further be verified.

a) b)

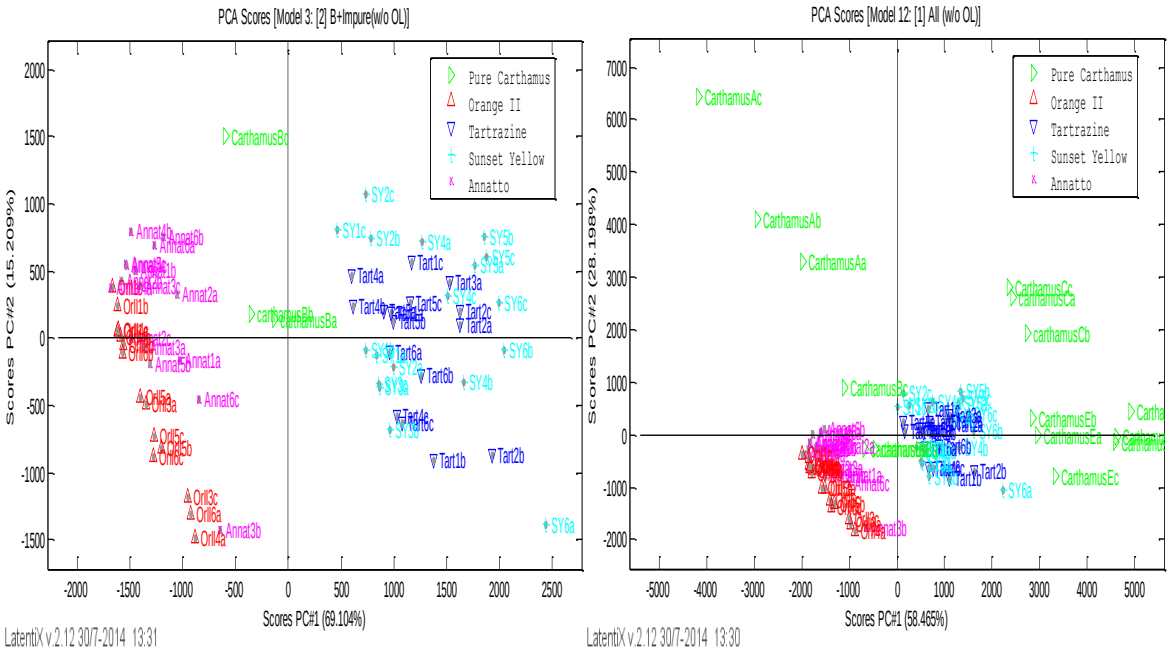


Figure 4.29 PCA Scores plot for the fluorescence data; a) including only the impure samples and the Carthamus B sample; b) including all the impure samples and all different types of pure Carthamus (Pre-processing of raw data: Mean Centering; Dilution: 1000 ppm).

QUANTITATIVE ANALYSIS

The same way as for UV-Visible data, a quantitative analysis of the fluorescence data was made. Hence, four PLS models were created for each unwanted additive. The data was pre-processed with mean centering and the models obtained were cross-validated. The PLS scores plots for the four unwanted additives are presented in Figure 4.30 and Figure 4.31. These figures also display the plot showing the RMSE for calibration and for validation according to the number of latent variables for the each corresponding color additive.

By analyzing the figures below, it can be observed that there is a linear trend for all the four additives according to its concentration. However, the tendency for RMSECV according to the number of latent variable is not favorable, which indicate that, after all, the samples do not behave in a linear trend. For that reason, the creation of a model able to quantify the additives becomes a quite difficult task.

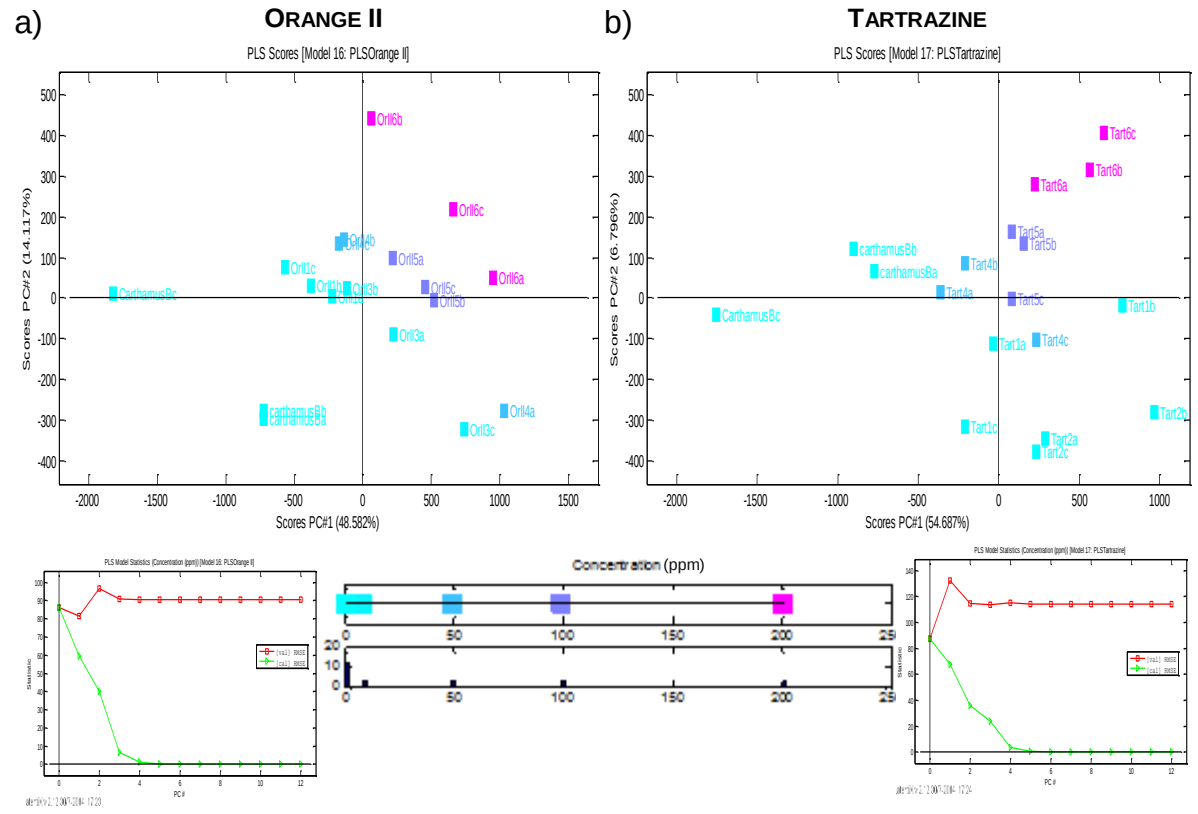


Figure 4.30 PLS Scores plot colored according to the concentration (in ppm) of the unwanted additive and the corresponding RMSE plot showing the RMSEC (green line) and RMSECV (red line) according to the number of LVs for the fluorescence data; a) for the Carthamus containing Orange II; b) for the Carthamus containing Tartrazine.

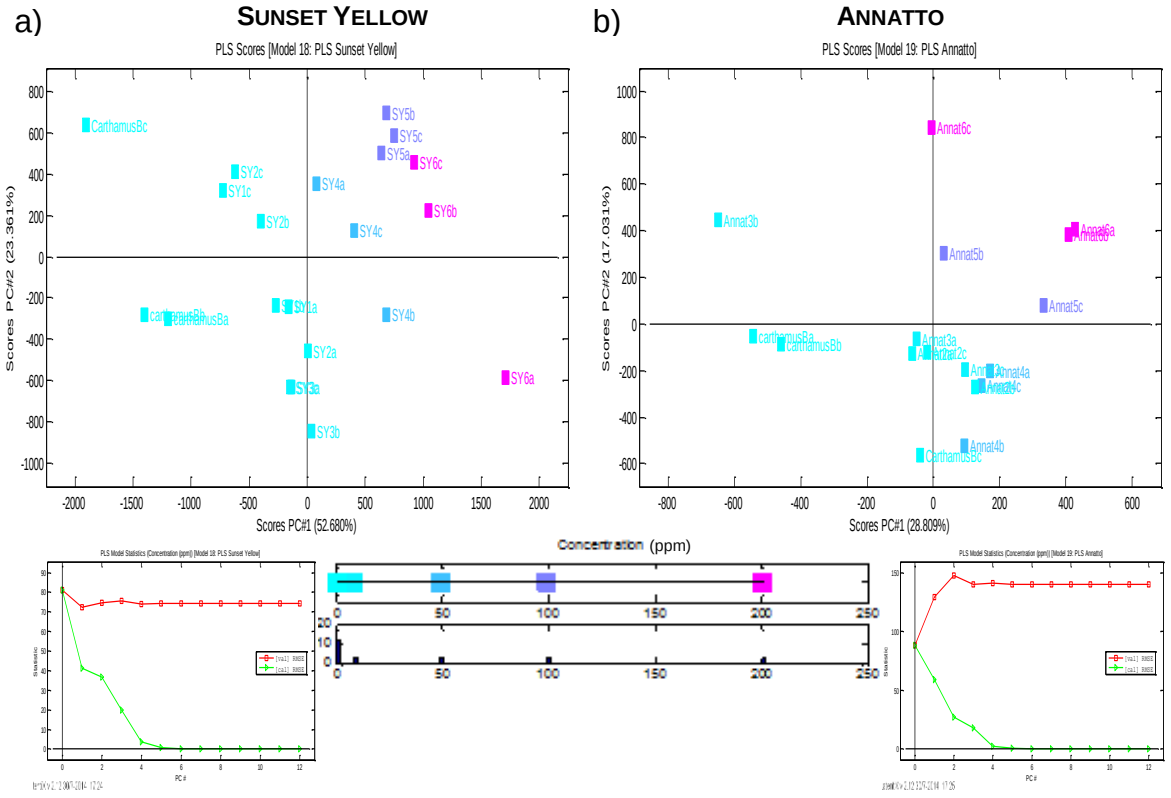


Figure 4.31 PLS Scores plot colored according to the concentration (in ppm) of the unwanted additive and the corresponding RMSE plot showing the RMSEC (green line) and RMSECV (red line) according to the number of

LVs for the fluorescence data; a) for the Carthamus containing Sunset Yellow; b) for the Carthamus containing Annatto.

QUALITATIVE ANALYSIS

In order to develop a model that is able to distinguish pure *Carthamus* samples from the impure ones, PLS-DA classification method was used, the same way as it has been previously performed with UV-Visible data. The scores plot and the RMSE for calibration and for validation according to the number of latent variables plot are presented in Figure 4.32 and Figure 4.33, respectively.

Considering the Figure 4.32, it can be noted the separation between pure and impure samples, however this distinction between the two classes is not linear, since two distinct groups of pure *Carthamus* are evident. And, for that reason, the classification model achieved through PLS-DA is not adequate and its classification efficiency is not completely satisfactory, as shown in the confusion matrix presented in Table 4.3.

Considering Figure 4.33, it should be noted that the tendency for the RMSECV does not follows the tendency for the RMSE for calibration, which, once again, indicates that there is no linear trend in this data.

In fact, by looking at the two groups for the pure samples (marked in green in Figure 4.32) in a different perspective, two different linear problems can be analyzed. Actually, this is a typical and not so common case, where PLS-DA does not work properly as a classification method. Therefore, a non-linear method should be experimented.

It is quite difficult to perceive the reason for this non-linearity and how the model actually performs as more data would be needed. In order to get more data, more varieties of *Carthamus* samples should have been used in the screening phase of the experimental procedure, instead of using just *Carthamus B.* Unfortunately, there was no time to for this task in the current project.

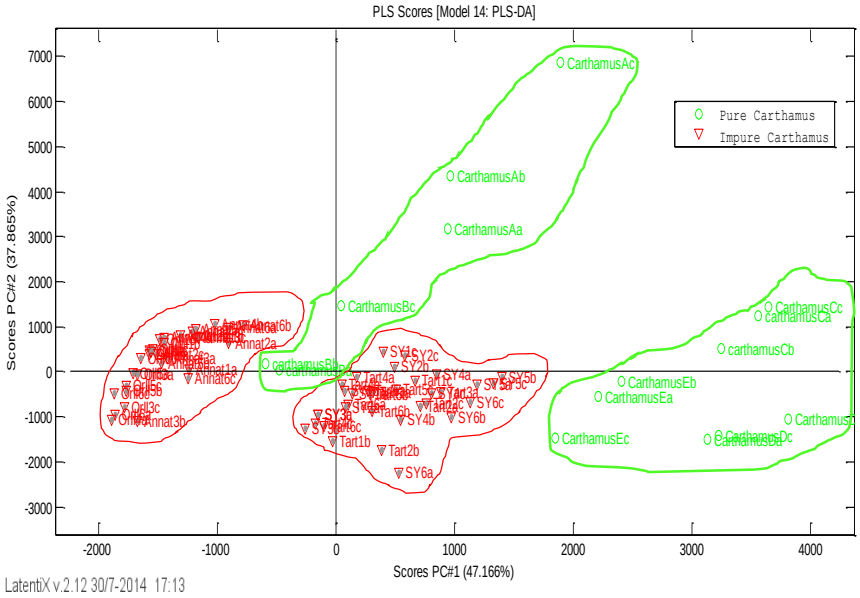


Figure 4.32 Scores plot for the data from the fluorescence data colored according to the two categories: Pure and Impure Carthamus (Pre-processing of raw data: Mean Centering; Dilution: 1000 ppm).

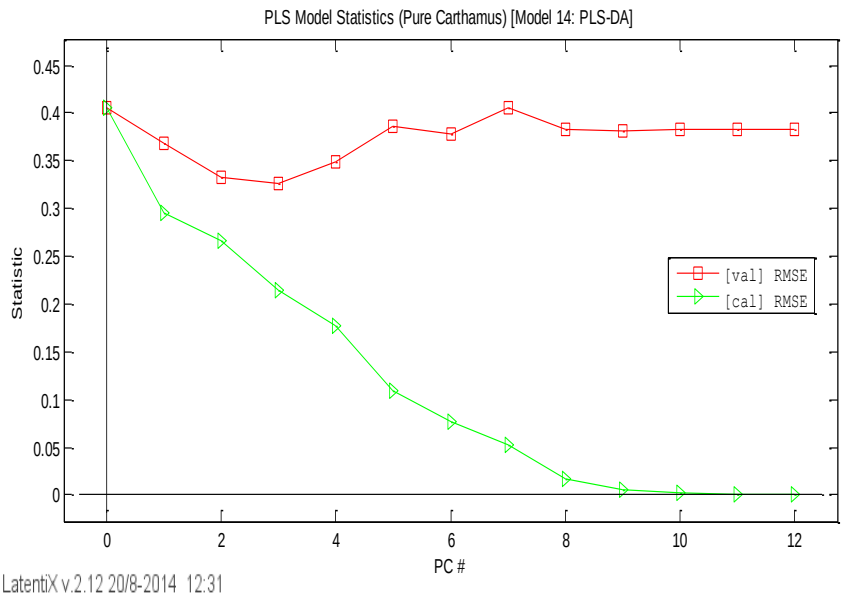


Figure 4.33 RMSE for Calibration (green line) and Cross-Validation (red line) intended for the Pure Carthamus class according to the number of latent variables used for the fluorescence data.

Table 4.3 PLS-DA classification confusion matrix for the fluorescence data.

		Pure	**Impure**
Predicted as	Pure	**9**	0
	Impure	6	**65**

Based on the fact that this data shows a non-linear trend, it is relevant to see how a non-linear classification method, such as k-Nearest Neighbors (k-NN), performs. Therefore, using the PLS toolbox in MATLAB, k-NN was implemented on the fluorescence data. The model was cross-validated with leave-one-out method but taking into account the samples replicates, inserting the three replicates together in the same cross-validation segment. In this case, the pre-processing method used was auto scaling. By using this technique the same weight is given to each variable/wavelength, which helps to focus k-NN on the informational part of the spectra.

By analyzing Figure 4.34, displaying the RMSEC and RMSECV against the number of k neighbors it is observed that the most appropriate number of neighbors is 3, since it has the lowest RMSECV, considering that only odd numbers are used in this cases of classification with two classes, to avoid ties.

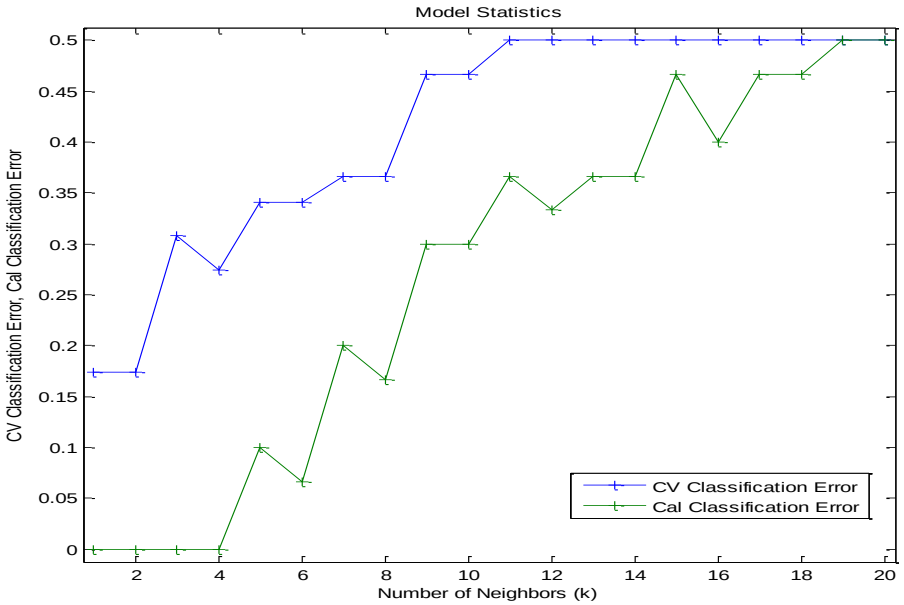


Figure 4.34 RMSE for calibration (green line) and for cross-validation (blue line) against the number of neighbors (Pre-processing of raw data: Auto scaling; Dilution: 1000 ppm).

In Figure 4.35, a plot showing the misclassified samples is presented, which indicates that only 3 pure *Carthamus* samples and 1 replicate of an impure sample are misclassified, however this is not the results for the cross-validated model. In Table 4.4, the confusion matrix for the k-NN model after cross-validation is shown.

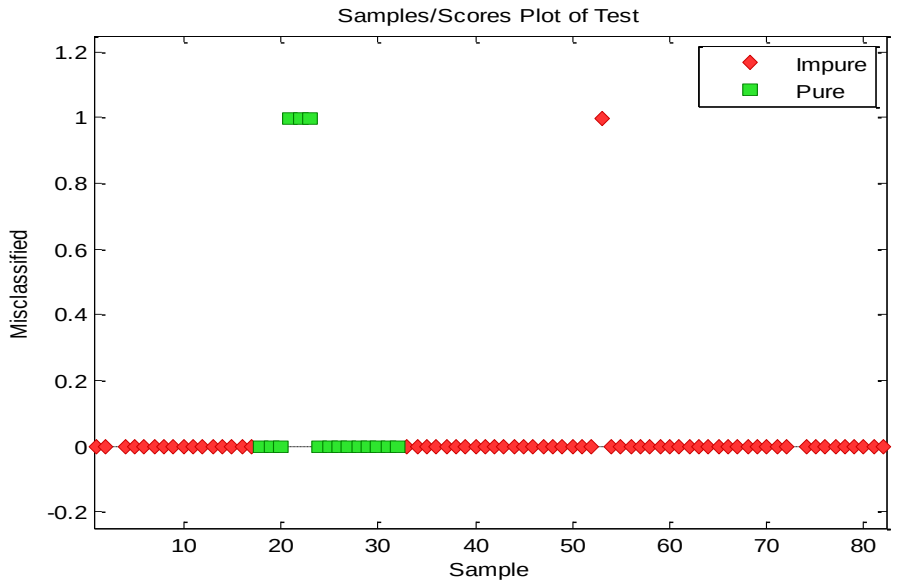


Figure 4.35 Misclassified samples for the two classes: Pure (green) and Impure (red) Carthamus.

Table 4.4 k-NN classification confusion matrix for the fluorescence data.

		Pure	**Impure**
Predicted as	Pure	**6**	1
	Impure	9	**64**

As can be seen in the confusion matrix (Table 4.4), the classification performance of k-NN is not entirely satisfactory, since this model misclassifies 9 pure *Carthamus* samples as impure and 1 impure sample (one replicate of *Carthamus* containing 1.6 mg of Sunset Yellow per kg of pure *Carthamus*) as pure, an inferior result to the one obtained with PLS-DA. Evidently, this is due to the closeness between these misclassified impure and pure samples to samples of the contrary category, as observed in Figure 4.32.

Even though *k*-NN method is not working properly in terms of classification, in Figure 4.32 it was possible to observe a difference between the spatial arrangement of the pure samples and the impure ones, which demonstrates that the optimized fluorescence data is valuable. However, to be completely certain, a more extensive data set would be required, which could be achieved by using more pure samples and more mixtures prepared with other types of pure Carthamus (rather than using only Carthamus B). Additionally, the fact that this data has an unbalanced density of samples for each category (the pure class has a much smaller number of samples than the impure class) might be an obstacle for the implementation of the *k*-NN method in this data, since the unbalanced data issue is considered one of the limitations of this method, according to [52].

Therefore, it is believed that the results do not relate to the classification method used, but rather with the dimension of this data and the spectroscopic method used.

For that reason, it is important to verify the detection ability of the fluorescence instrument. Therefore, a detection test was made in which 160 ppb of each of the three synthetic color additives (Orange II, Tartrazine and Sunset Yellow) were added to water in order to simulate the same quantity previously added to the *Carthamus* natural product and, consequently, verify if the detection of this amount of additive was possible. The resulting fluorescence landscape is presented in Figure 4.36. In this case, there is no need to choose an excitation range of the landscape since it is possible to evaluate the results from the landscape itself.

As can be observed in this figure, the detection of the additives in water is quite weak. The diagonal line observed is a Raman scatter line, while the horizontal/vertical lines are artifacts associated with the instrument setup.

From this trial it is possible to conclude that the detection of such low quantity of unwanted additive using a fluorescence instrument is not very strong. Unfortunately, this indicates that even with the optimized setting it is not possible to obtain the desired results for this project.

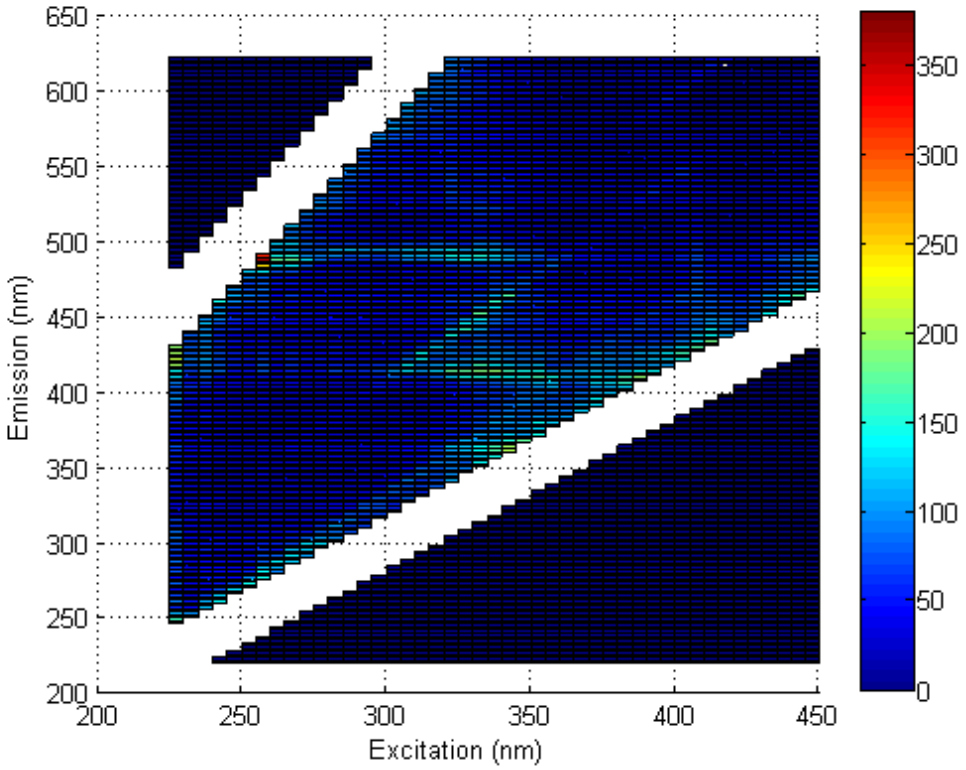


Figure 4.36 Fluorescence landscape for the detection test made for three water samples containing 160 ppb of each synthetic additive.

5. CONCLUSIONS AND FUTURE PROSPECTS

This study focused on using multivariate data analysis on spectral data with the aim of creating quantitative and qualitative methods/models that could be used to trace adulterations, concerning the addition of undesirable color agents - Orange II, Tartrazine, Sunset Yellow and Annatto - in the natural food coloring product – *Carthamus* Yellow – commercialized by the food industry.

In recent years, the food industry has been investing much more in its control systems due to the increasing demand for higher quality products, concerning food safety, required not only by the regulatory agencies but also by the consumers themselves. This demand for superior quality products also stands as a differentiation factor in such competitive business as the food industry. Therefore, this research is of great interest since there have been many reported cases of contaminations on the raw materials coming from the suppliers, which jeopardizes the quality of the final product.

Initially, the spectroscopic instruments used to acquire the data considered as suitable for the present work were selected – UV-Visible, since the instrument was already available at Chr.Hansen laboratories; Fluorescence, since it is a very sensitive instrument; and NIR as a commonly used instrument for analysis of food stuff.

After the acquisition of the spectral data, the first step was to apply the exploratory analysis method, PCA (Section 4.4.1).This first analysis indicated that NIR data would not be very useful to achieve the objectives of the project since it has the limitation of not detecting low concentrations solutions, as the ones used in this work. For that reason, NIR data was discarded from further analysis.

By comparing the results for the UV-Visible and Fluorescence data, UV-Visible showed slightly superior results and thus this was the first data used in further analysis (Section 4.4.2).

Firstly, a quantitative analysis was executed in which PLS regression was performed with the objective of creating four quantitative models for each of the unwanted additives. The absence of a defined trend in the resulting PLS scores plots and the high values of the RMSE for the four additives led to the conclusion that the UV-Vis data for quantitative purposes is not suitable. A possible solution would be to consider a wider range of concentrations, however the main problem found here relates to the quantity of unwanted additives in the samples. The fact that in UV-visible spectroscopy there is the need of diluting the samples makes the quantity of unwanted additives very small.

Secondly, a qualitative analysis was accomplished, in which two classification methods were experimented – PLS-DA and SIMCA. On one hand, the PLS-DA model created showed a good classification capability that indicates that, even though this model has its limitations, it could be used. On the other hand, SIMCA presented inferior results which laid aside this model as a tool for the classification of *Carthamus* samples.

Considering that the results obtained with the UV-Visible data were not entirely satisfactory a linearity and detection test were performed. These tests showed that the concentration of the additives is linear, however the detection test confirmed that the low quantity of unwanted additives in the *Carthamus* Yellow is hardly detected by the UV-Vis spectrophotometer, which was identified as the main obstacle encountered in this study.

Subsequently, a further evaluation of the fluorescence data was made (Section 4.4.3). This data was analyzed with a similar approach as the UV-Visible data analysis. Here again, by implementing PLS

regression to create a quantitative model it was concluded that this model was not able to fulfil the objective of quantifying the unwanted additives, for the same reasons as for UV-Visible data.

Finally, two classification methods - PLS-DA and k-NN – were applied to the fluorescence data. The result for the first method used, PLS-DA, indicated that it would be better to use a different method considering the non-linearity verified in the data and, hence, k-NN was used as it suits better for non-linear distributions of the data in the scatter plot. Unfortunately, the model acquired with k-NN, comparing to the one obtain with PLS-DA, did not showed a superior result in terms of performance of classification of the samples as *Carthamus* Pure or Impure.

The fact that k-NN was identified as a more suitable method for this situation does not translate in superior results since it is believed that the issues here addressed relate to the spectroscopic instruments used and, consequently, the data acquired. This fact was also supported by the detection test made for fluorescence data that demonstrated, once again, that the detection of such low quantities of additives is very complicated. Therefore, it is possible to conclude that the outcome from the Fluorescence spectroscopy did not overcome what was obtained with UV-Visible.

To briefly summarize, it was observed that none of the spectroscopic methods allowed the creation of models that could quantify the undesired additives present in the natural product *Carthamus* Yellow. However, this was not the most important objective of this study and therefore, the main focus was elaborate a model that could identify if a sample is in pure form or was adulterated with other compounds. Thus, for classification purposes, the best results were obtained through the PLS-DA method using the UV-Visible data and, therefore, this method and model can be used by Chr.Hansen as a screening test for the *Carthamus* samples. However, by analyzing the optimized Fluorescence spectra, it was observed that a more accurate classification model could be created if a more extensive data set was used.

To better understand which is behind the obstacles encountered in this study it would be interesting to acquire a more complete data set, which could be achieved by executing additional screening trials for all the five different *Carthamus* provided by Chr.Hansen, rather than using only one type of *Carthamus*. This would create more data and, consequently, it would be possible to create more comprehensive models. Due to the time limitations of this work, it was not possible to do more laboratory work.

Even though the outcome of this project was not as successful as intended, it is important to note that not only this was the first study on this topic involving the *Carthamus* yellow product, but also the findings here achieved consist of good and useful material for further works on related subjects.

Further work could be done to establish whether the acquirement of more data, as previously suggest, could overcome the issues faced in this project. The Fluorescence spectroscopy as highly sensitive method is indicated as an appropriated method in further researches. Hence, it would be particularly interesting to investigate an adequate measurement mode in which the samples could be measured in pure form, rather than diluting them in water since this drastically reduces the quantity of additives present in the mixtures.

6. REFERENCES

[1] REGULATION (EC) No 1333/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 2008 on food additives, OJ L354/16, 2008.

[2] COMMISSION REGULATION (EU) No 231/2012 of 9 March 2012 laying down specifications for food additives listed in Annexes II and III to Regulation (EC) No 1333/2008 of the European Parliament and of the Council, OJ L83/1, 2012.

[3] Chr.Hansen, "Natural Colors: Coloring Foodstuffs." [Online]. Available: http://www.chr-hansen.com/products/product-areas/natural-colors/our-product-groups/fruitmaxr.html. [Accessed: 23-Jul-2014].

[4] R. Guidetti, R. Beghi, and V. Giovenzana, "Chemometrics in Food Technology," in *Chemometrics in Practical Applications*, InTech, 2012, pp. 217–252.

[5] European Comission, "RASFF notification on Orange II dye," 2009. [Online]. Available: http://www.spsvietnam.gov.vn/Lists/Ti liu/Attachments/736/phu luc 615-cl2.pdf. [Accessed: 04-Jun-2014].

[6] JECFA, "Carthamus Yellow," 2002. [Online]. Available: http://www.fao.org/ag/agn/jecfa-additives/specs/Monograph1/Additive-119.pdf. [Accessed: 15-Jun-2014].

[7] C.-C. Wang, C.-S. Choy, Y.-H. Liu, K.-P. Cheah, J.-S. Li, J. T.-J. Wang, W.-Y. Yu, C.-W. Lin, H.-W. Cheng, and C.-M. Hu, "Protective effect of dried safflower petal aqueous extract and its main constituent, carthamus yellow, against lipopolysaccharide-induced inflammation in RAW264.7 macrophages.," *J. Sci. Food Agric.*, vol. 91, no. 2, pp. 218–25, Jan. 2010.

[8] J.-M. Yoon, M.-H. Cho, J.-E. Park, Y.-H. Kim, T.-R. Hahn, and Y.-S. Paik, "Thermal stability of the pigments hydroxysafflor yellow A, safflor yellow B, and precarthamin from safflower (*Carthamus tinctorius*)," *J. Food Sci.*, vol. 68, no. 3, 2003.

[9] Z. Ekin, "Resurgence of safflower (*Carthamus tinctorius* L.) utilization: a global view," *J. Agron.*, no. 4, pp. 83–87, 2005.

[10] REGULATION (EC) No 178/2002 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety, 2002.

[11] REGULATION (EC) No 1334/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 2008 on flavourings and certain food ingredients with flavouring properties for use in and on foods and amending Council Regulation (EEC) No 1601/91, Regulations (EC) No 2232/96 and (EC) No 110/2008 and Directive 2000/13/EC, 2008.

[12] Chr. Hansen, "Raw material inspection plan – Carthamus Yellow," 2014.

[13] V. Emongor, "Safflower (*Carthamus tinctorius* L.) the underutilized and neglected crop: A review," *Asian J. Plant Sci*, 2010.

[14] R. M. Christie, "Azo dyes and pigments," in *Colour Chemistry*, The Royal Society of Chemistry, 2001, pp. 2001, 45–68.

[15] EFSA Review, "Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food on a request from the Commission to Review the toxicology of a number of dyes illegally present in food in the EU," *Efsa J.*, no. 263, pp. 1–71, 2005.

[16] Sigma-Aldrich, “Orange II sodium salt (CAS: 633-96-5) Product Specification.” [Online]. Available: http://www.sigmaaldrich.com/catalog/product/aldrich/195235?lang=en&region=DK. [Accessed: 01-Jul-1014].

[17] IARC, “IARC Monographs - Classification List,” 2014. [Online]. Available: http://monographs.iarc.fr/ENG/Classification/ClassificationsCASOrder.pdf. [Accessed: 05-Jul-2014].

[18] EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS), “Scientific Opinion on the re-evaluation Tartrazine (E 102),” *EFSA J.*, vol. 7, no. 11, 2009.

[19] Sigma-Aldrich, “Tartrazine (CAS: 1934-21-0) Product Specification.” [Online]. Available: http://www.sigmaaldrich.com/catalog/product/sigma/t0388?lang=en&region=DK. [Accessed: 01-Jul-2014].

[20] EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS), “Scientific Opinion on the re-evaluation of Sunset Yellow FCF (E 110) as a food additive,” *EFSA J.*, vol. 7, no. 11, 2009.

[21] “Sunset Yellow FCF (CAS: 2783-94-0) Product Specification.” [Online]. Available: http://www.sigmaaldrich.com/catalog/product/aldrich/465224?lang=pt&region=PT. [Accessed: 01-Jul-2014].

[22] Eurofins, “WEJ Contaminants: Sudan dyes and other illegal dyes.” [Online]. Available: http://www.eurofins.de/media/2717988/sudan_eng.pdf. [Accessed: 10-Jul-2014].

[23] James Smith, “Annatto Extracts - Chemical and Technical Assessment,” PhD Thesis, 2006.

[24] É. Dufour, “Principles of Infrared spectroscopy,” in *Infrared Spectroscopy for Food Quality Analysis and Control*, 2009, pp. 3–28.

[25] D. A. Skoog and D. M. West, *Principles of Instrumental Analysis*. 1980.

[26] K. Wiberg, “Multivariate Spectroscopic Methods for the Analysis of Solutions,” 2004.

[27] T. Owen, *Fundamentals of modern UV-visible spectroscopy*. 2000.

[28] D. L. Pavia, G. M. L. G. S. Kriz, and J. R. Vyvyan, *Introduction to spectroscopy*, 4th ed. 2009.

[29] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*. Chichester, UK: John Wiley & Sons, Ltd, 2004.

[30] G. Gauglitz and T. Vo-Dinh, *Handbook of spectroscopy*. 2006.

[31] D. Baunsgaard, “Analysis of Color Impurities in Sugar Processing using Fluorescence Spectroscopy and Chemometrics,” Ph.D. Thesis, Spectroscopy and Chemometrics Group - Food Science Department, University of Copenhagen, 2000.

[32] J. Lackowicz, *Principles of Fluorescence Spectroscopy*, 3rd ed. Springer, 2006.

[33] B. Valeur, *Molecular Fluorescence: Principles and Applications*, 3rd ed., Wiley, 2001.

[34] J. R. Albani, *Principles and Applications of Fluorescence Spectroscopy*. Blackwell, 2007.

[35] A. J. Lawaetz, “Fluorescence Spectroscopy and Chemometrics - Applied ni Cancer Diagnostics and Metabolomics,” Ph.D. Thesis, Spectroscopy and Chemometrics Group - Food Science Department, University of Copenhagen, 2011.

[36] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent, “A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies,” *J. Pharm. Biomed. Anal.*, vol. 44, no. 3, pp. 683–700, Jul. 2007.

[37] D. A. Burns and E. W. Ciurczak, *Handbook of Near-Infrared Analysis*, 3rd ed. CRC Press, 2008.

[38] H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise, *Near-Infrared Spectroscopy: Principles, Instruments, Applications*, Wiley-vch, 2002.

[39] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, P. J. L. S. De Jong, and J. Smeyers-Verbeke, “Handbook of Chemometrics and Qualimetrics (Part A),” in *Data Handling in Science and Technology*, Elsevier, Ed. 1998.

[40] K. Varmuza, *Chemometrics in Practical Applications*. InTech, 2012.

[41] M. L. Vigni, C. Durante, and M. Cocchi, “Exploratory Data Analysis,” in *Chemometrics in Food Chemistry*, Elsevier, pp. 55–126, 2013.

[42] M. B. Romía and M. A. Bernàrdez, “Multivariate Calibration for Quantitative Analysis,” in *Infrared Spectroscopy for Food Quality Analysis and Control*, pp. 51–82, Elsevier, 2009.

[43] T. Naes, T. Isaksson, T. Fearn, and T. Davies, *A user friendly guide to Multivariate Calibration and Classification*. NIR Publications, 2002.

[44] M. J. Adams, *Chemometrics in Analytical Spectroscopy*. The Royal Society of Chemistry, 1995.

[45] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, P. J. L. S. De Jong, and J. Smeyers-Verbeke, “Handbook of Chemometrics and Qualimetrics (Part B),” in *Data Handling in Science and Technology*, Elsevier, 1998.

[46] F. Westad, M. Bevilacqua, and F. Marini, “Regression,” in *Chemometrics in Food Chemistry*, vol. 28, Elsevier, pp. 127–170, 2013.

[47] M. Andersson, “A comparison of nine PLS1 algorithms,” *J. Chemom.*, vol. 23, no. 10, pp. 518–529, Oct. 2009.

[48] S. Wold, H. Martens, and H. Wold, “The cultivariate calibration problem in Chemistry solved by the PLS method,” *Lect. Notes Math.*, vol. 973, pp. 286–293, 1983.

[49] S. de Jong, “SIMPLS: An alternative approach to partial least squares regression,” *hemometrics Intell. Lab. Syst.*, vol. 18, pp. 251–263, 1993.

[50] N. M. Faber and J. Ferré, “On the numerical stability of two widely used PLS algorithms,” *J. Chemom.*, vol. 22, no. 2, pp. 101–105, Feb. 2008.

[51] Å. Rinnan, F. Van Den Berg, and S. B. Engelsen, “Review of the most common pre-processing techniques for near-infrared spectra,” *Trends Anal. Chem.*, vol. 28, no. 10, pp. 1201–1222, Nov. 2009.

[52] M. Bevilacqua, R. Bucci, A. D. Magrı, A. L. Magrı, R. Nescatelli, and F. Marini, “Classification and Class-Modelling,” in *Chemometrics in Food Chemistry*, Elsevier, pp. 171–233, 2013.

[53] D. Ballabio and R. Todeschini, “Multivariate Classification for Qualitative analysis,” in *Infrared Spectroscopy for Food Quality Analysis and Control*, pp. 83–104, Elsevier, 2009.

[54] “LatentiX Version 2.12 - Multivariate Analysis Software.” [Online]. Available: http://www.latentix.com/layout.asp?index. [Accessed: 17-Jun-2014].

[55] M. G. Lima, I. M. Raimundo, and M. F. Pimentel, “Improving the detection limits of near infrared spectroscopy in the determination of aromatic hydrocarbons in water employing a silicone sensing phase,” vol. 125, pp. 229–233, 2007.

APPENDICES

APPENDIX I

SETTINGS FOR THE SPECTROSCOPIC METHODS

UV-Visible Spectrophotometer

Table 1 Settings used in the UV-Vis Spectrophotometer for all the Carthamus samples.

Data mode	Absorbance
Start wavelength	1056 nm
End wavelength	300 nm
Bandwidth	2 nm
Integration time	0.20 sec
Data interval	2.00 nm
Scan speed	600 nm/min

Table 2 Settings used in the UV-Vis Spectrophotometer for the unwanted additives samples.

Data mode	Absorbance
Start wavelength	1056 nm
End wavelength	200 nm
Bandwidth	2 nm
Integration time	0.20 sec
Data interval	2 nm
Scan speed	600 nm/min

Fluorescence Spectrophotometer

Table 3 Settings used in the Fluorescence Spectrophotometer for the diluted samples (clear solutions).

	Start	**Step**	**End**	**Slit**
Excitation (nm)	250	5	350	10
Emission (nm)	250	2	650	10
Scan control	Fastest –Step:2 nm			
Scan rate	9600 nm/min			
Detector voltage	Medium - 600 V			

Table 4 Settings used in the Fluorescence Spectrophotometer for the opaque samples, the Carthamus in pure form or mixed with unwanted additives.

	Start	Step	End	Slit
Excitation (nm)	250	5	350	20
Emission (nm)	300	2	650	20
Scan control	Fastest –Step:2 nm			
Scan rate	9600 nm/min			
Detector voltage	Medium - 600 V			

Table 5 Settings used in the Fluorescence Spectrophotometer for solids (powder samples: Orange II, Tartrazine and Sunset Yellow)

	Start	Step	End	Slit
Excitation (nm)	250	10	500	10
Emission (nm)	250	2	800	10
Scan control	Fastest –Step:2 nm			
Scan rate	9600 nm/min			
Detector voltage	Medium - 600 V			

Near-Infrared Spectrophotometer

Table 6 Settings used in the Near-Infrared Spectrophotometer.

Spectral Range	Start	10000 cm^{-1}
	End	4000 cm^{-1}
Initial Delay		0 sec
Number of Scans		64
Data type		Absorbance
Resolution		8 cm^{-1}

APPENDIX II

MATLAB scripts

This appendix displays the MATLAB scripts (kindly provided by Prof. Åsmund Rinnan) that served as a basis for the conversion of the files from the three instruments to Matlab files.

UV-Visible

```
d = dir('*.csv');
d = cellstr( char( d.name) );
d %to confirm if it is the right directory
X = readuv( d);
X %to see the details of the matrix
plot( X.Axis, X.Spec)
DM = [X.Spec; X.Axis'];
SN = char( [X.SampleID; {'Wavelength (nm)'}]);
VL = num2str( X.Axis);
save docname DM SN VL
```

Fluorescence

```
X = readeem;

DM = X.Unfold;
SN = char( X.SampleID);
VL = X.ULab;
save filename DM SN VL
figure; surf( X.Ex, X.Em, squeeze( X.EEM( 1, :, :) ) )

%cutting
[xnew, emnew, exnew] = flucut( X.EEM, X.Em, X.Ex, [-15 15], [-20 20]);
figure; surf( exnew, emnew, squeeze( xnew( 1, :, :) ) )

opt = eem2ltx;
opt.CutLower = [-15 NaN];
opt.CutUpper = [-20 NaN];
[DM, SN, VL] = eem2ltx( X, opt);
save cutfilename DM SN VL
```

NIR

```
d = dir('*.spc');
d = cellstr( char( d.name) );
d %to confirm if it is the right directory
X = readspc( d);
X %to see the details of the matrix
plot( X.Axis, X.Spec)
for cf = 1:length( X)
    DM = [X(cf).Spec; X(cf).Axis(:)'];
    SN = char( [X(cf).SampleID; {'Wavelength (nm)'}]);
    VL = num2str( X(cf).Axis(:));
    save( ['docname' num2str(cf)], 'DM', 'SN', 'VL')
end
```

APPENDIX III

RESULTS FOR THE REDUCED SPECTRAL RANGE

This appendix shows the PCA score plot obtained for the reduced spectral range (320-420 nm) for the UV-Visible data.

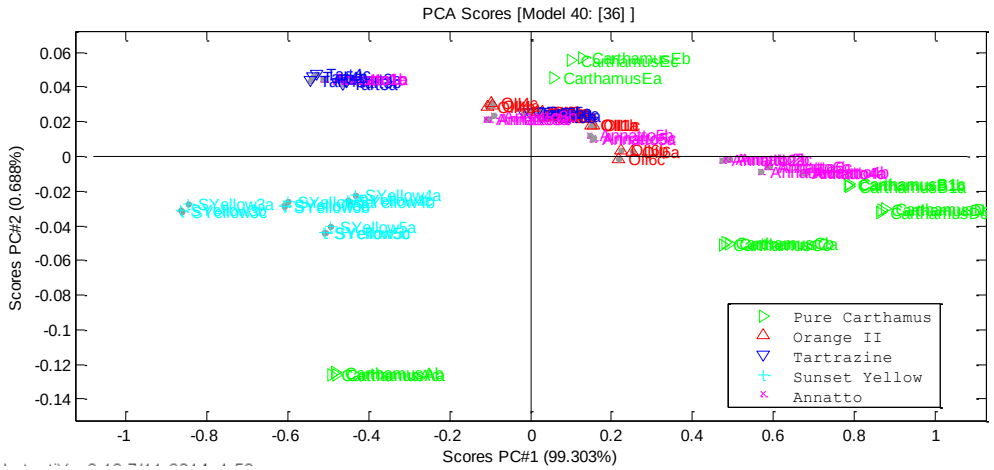


Figure 1 PCA score plot using a reduced spectral range from 320 nm to 420 nm for the UV-Visible data.

APPENDIX IV

RESULTS FOR THE UV-VISIBLE SPECTROSCOPY USING A HIGHER DILUTION IN WATER OF THE SAMPLES

This appendix contains the results obtained from the measurements made on the pure and impure *Carthamus* samples diluted in water to achieve a 10 ppb concentration. This dilution was made due to the fact that the detector might saturate in certain samples when a 1:1000 dilution was used. For the data acquired a multivariate analysis was also performed.

The raw spectra for the pure Carthamus samples is presented in Figure 2. The raw spectra for the Carthamus containing each of the undesired additives, Orange II, Tartrazine, Sunset Yellow and Annatto, are presented in Figure 3 a) and b), Figure 4 a) and b), respectively (colored according to the concentration of additive). The PCA scores plots containing all the samples and without the outliers are presented in Figure 5 a) and b), respectively. Additionally, Figure 6 and Figure 7 show the PLS score plot for each undesired additive (colored according to the concentration of additive). Finally, the results for the classification using PLS-DA are also presented in Table 7.

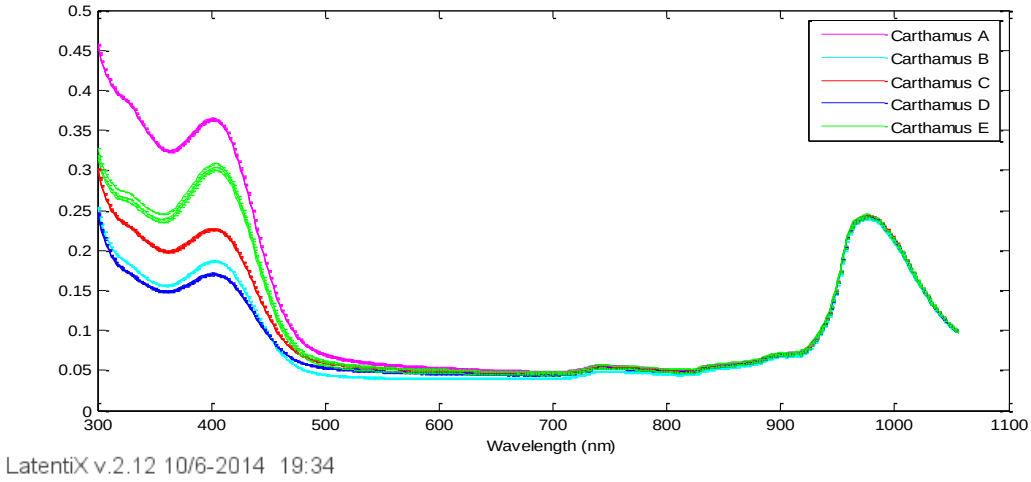


Figure 2 Raw spectra for the 5 different pure Carthamus samples diluted in a 10 ppb solution.

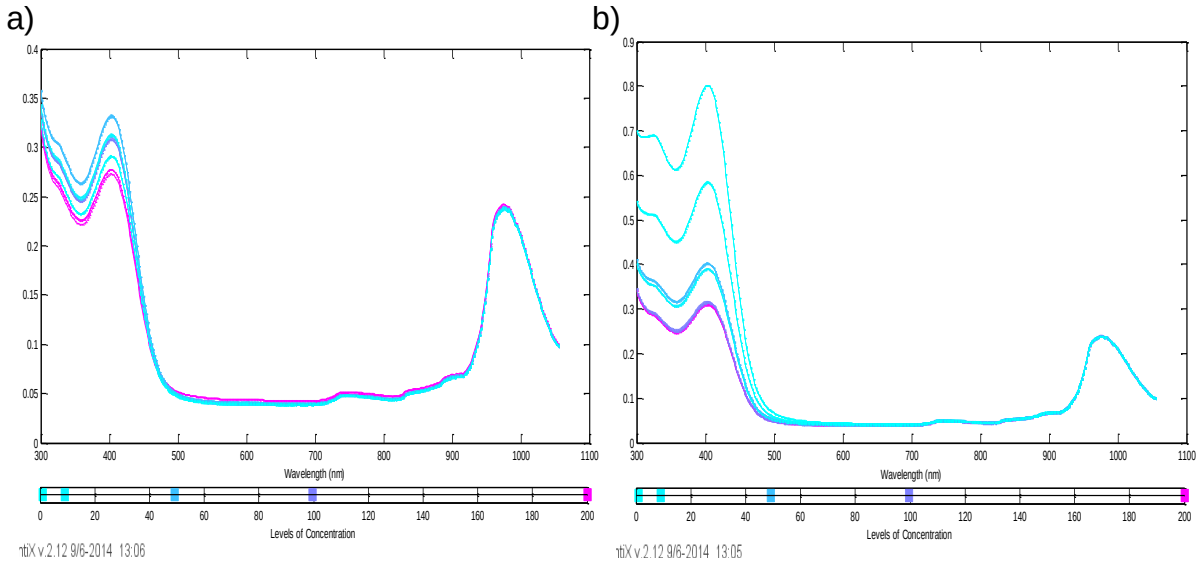


Figure 3 Raw spectra colored according to the concentration of the unwanted additives for the Carthamus samples containing: a) Orange II; b) Tartrazine.

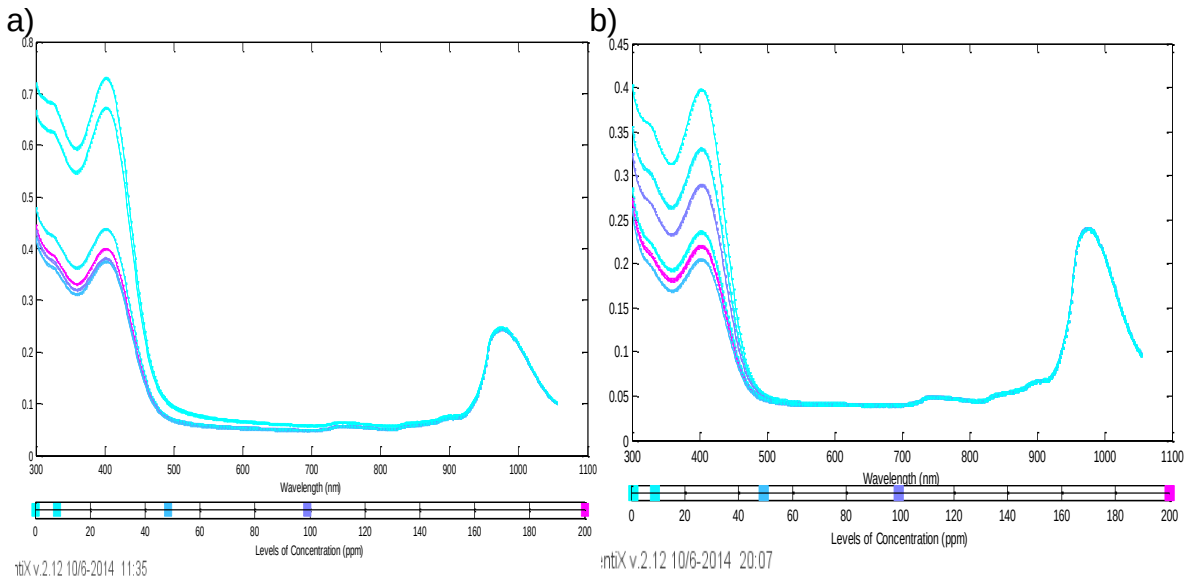


Figure 4 Raw spectra colored according to the concentration of the unwanted additives for the Carthamus samples containing: a) Sunset Yellow; b) Annatto.

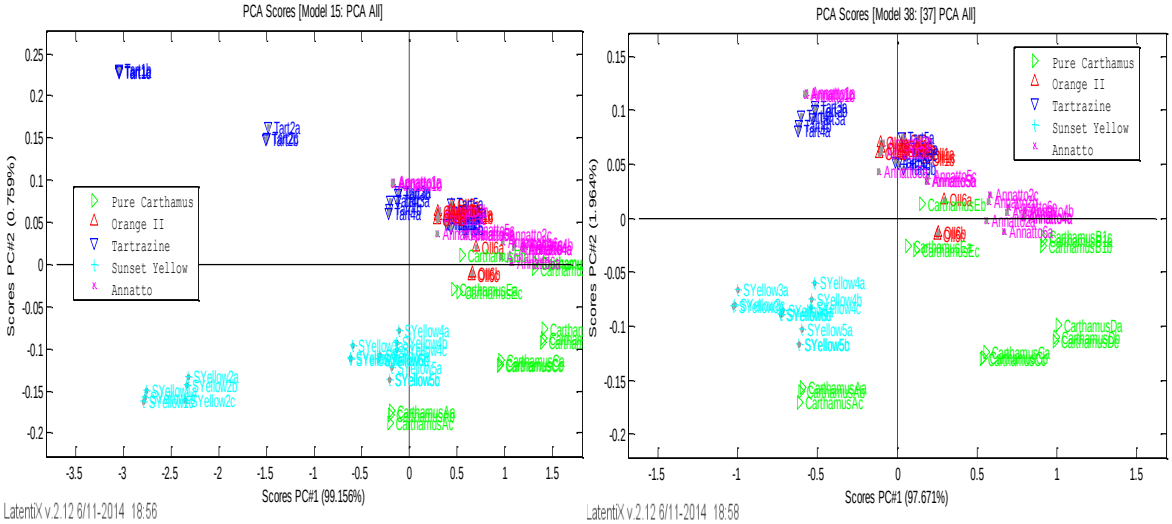


Figure 5 a) Score plot with all samples: Pure and Impure Carthamus. b) Score plot for all samples after outlier removal.

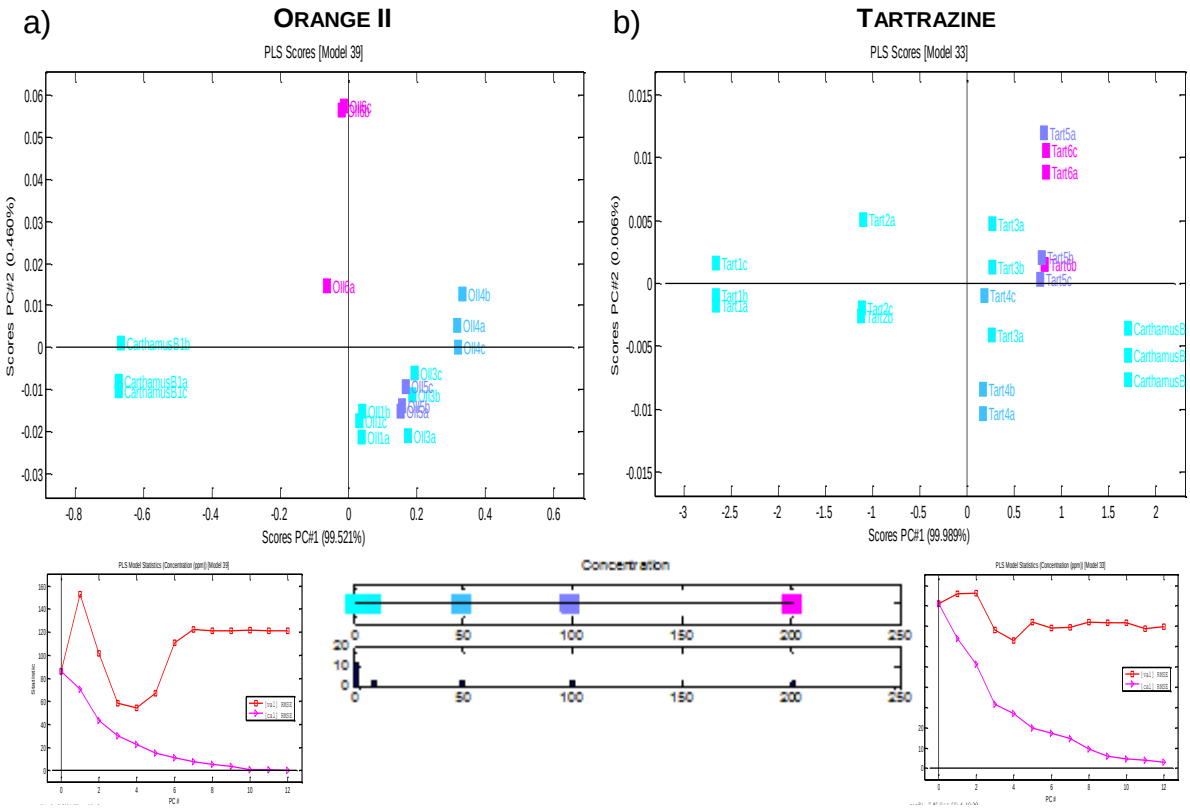


Figure 6 PLS Score plot colored according to the concentration of the unwanted additive and the corresponding RMSE plot showing the RMSEC (pink line) and RMSECV (red line) according to the number of LVs; a) for the unwanted additive Orange II; b) for the unwanted additive Tartrazine.

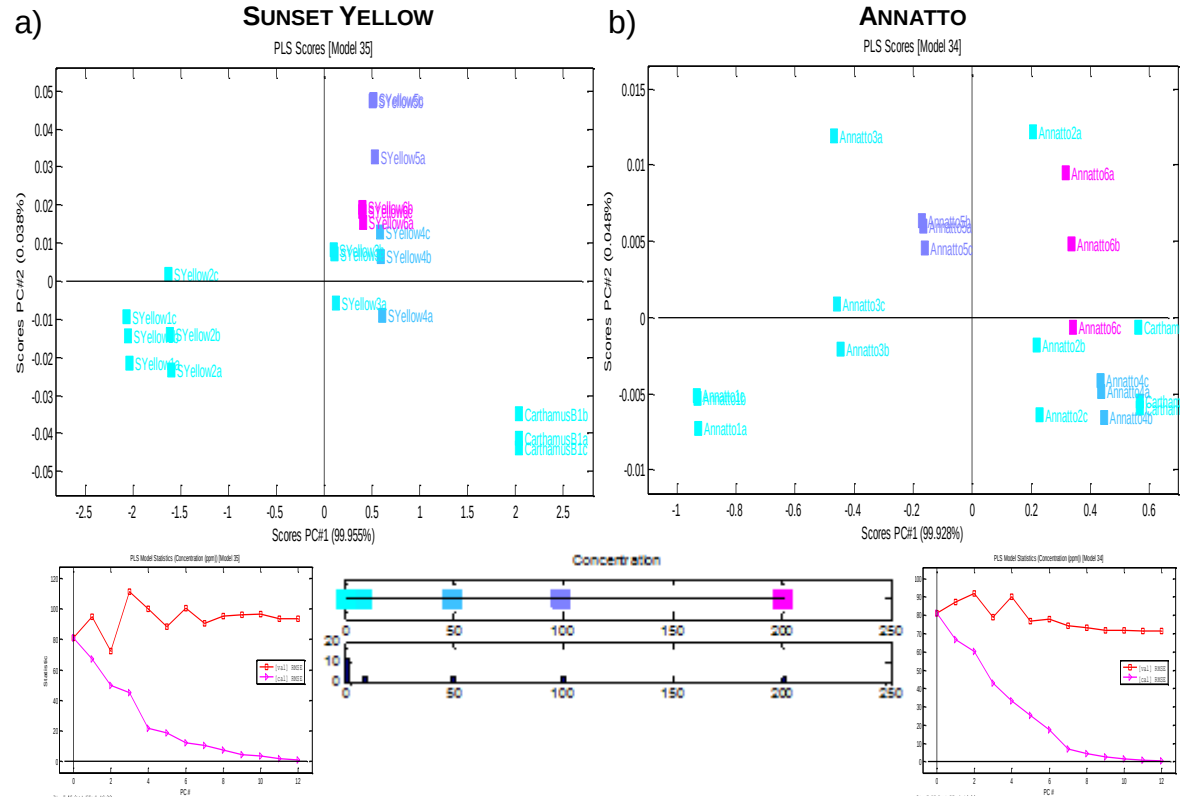


Figure 7 PLS Score plot colored according to the concentration of the unwanted additive and the corresponding RMSE plot showing the RMSEC (pink line) and RMSECV (red line) according to the number of LVs; a) for the unwanted additive Sunset Yellow; b) for the unwanted additive Annatto.

Table 7 PLS-DA Confusion matrix for the trial using the samples diluted in a 10 ppb water solution for the 2 categories: Pure and Impure Carthamus.

		Pure	Impure
Predicted as	Pure	6	0
	Impure	9	69

APPENDIX V

PLS-DA RESULTS FOR 5 CATEGORIES

In this appendix the confusion matrices for the PLS-DA perform using the 5 categories - Pure Carthamus and the four samples of impure Carthamus contaminated with five different illegal additives - Orange II, Tartrazine, Annatto and Sunset Yellow – are presented in Table 8 and Table 9 for the 1:1000 sample dilution and 10 ppb dilution in water, respectively.

Table 8 Confusion matrix for the trial using the samples diluted in a 1:1000 water solution for the 5 categories: Pure Carthamus and the four samples of impure Carthamus contaminated with five different illegal additives - Orange II, Tartrazine, Annatto and Sunset Yellow.

		Carthamus	Orange II	Tartrazine	Annatto	Sunset Yellow
Predicted as	Carthamus	57	0	0	0	0
	Orange II	0	0	0	10	0
	Tartrazine	0	0	15	0	3
	Annatto	0	15	0	0	11
	Sunset Yellow	0	0	0	6	1

Table 9 Confusion matrix for the trial using the samples diluted in a 10 ppb water solution for the 5 categories: Pure Carthamus and the four samples of impure Carthamus contaminated with five different illegal additives - Orange II, Tartrazine, Annatto and Sunset Yellow.

		Pure	Orange II	Tartrazine	Annatto	Sunset Yellow
Predicted As	Pure	7	13	18	0	0
	Orange II	0	0	0	0	0
	Tartrazine	0	0	0	0	0
	Annatto	1	2	0	18	0
	Sunset Yellow	7	0	0	0	18