Universitat Autònoma de Barcelona

Development of analytical methods based on Near Infrared Spectroscopy for monitoring of pharmaceutical and biotechnological processes and control of new psychoactive substances

Aira Yira Miró Vera

Universitat Autònoma de Barcelona

Development of analytical methods based on Near Infrared Spectroscopy for monitoring of pharmaceutical and biotechnological processes and control of new psychoactive substances

Aira Yira Miró Vera

Doctoral Thesis

PhD Program in Chemistry

Professor Manel Alcalà Bernàrdez, PhD, Director

Department of Chemistry

Faculty of Sciences

2019

Memòria presentada per aspirar al Grau de Doctor per

Aira Yira Miró Vera

Vist i plau

Manel Alcalà

Director: Professor Dr. Manel Alcalà Bernàrdez

Bellaterra, 17 de juliol del 2019

Acknowlegments

This research was funded by MINECO (Spain) through the project CTQ 2016-79696-P (AEI/FEDER, EU). I appreciate the academic support of Professor Manel Alcalà Bernàrdez. I am also grateful to Professor Francisco Valero and personnel of the Department of Chemical Engineering UAB, Professor Sergi Armenta and personnel of the Department of Analytical Chemistry University of Valencia for the kind provision of samples, instruments, as well as their valuable collaborative efforts.

ABSTRACT

The Near Infrared Spectroscopy (NIRS) is an analytical technique based on the interaction of electromagnetic radiation in the wavelength range 780-2500 nm and matter. The NIR spectrum can be considered as a "*fingerprint*" of each chemical compound or mixture of them, which contains absorption bands that are the result of overtones and combinations of the fundamental vibrations observed in the Mid-Infrared region. Additionally, NIRS of solids is sensible to the scattering effect caused by physical characteristics of the samples, therefore provides simultaneous sensitivity to chemical and physical changes of solids. Because of NIR spectra show broad and overlapped bands makes it necessary the use of Chemometrics, which implies the application of statistical and mathematical methods to such spectral data, for achieving the maximal extraction and collection of useful information from it. The general objective of this doctoral thesis is developing analytical methods based on NIRS for monitoring pharmaceutical and biotechnological processes and for the control of illicit drugs. The following conditions have been considered i) extended active pharmaceutical ingredient (API) concentration ranges during granulation and tableting using the process spectrum (PS), ii) on-site identification of new psychoactive substances (NPS) during police seizing procedures with hand-held and bench-top instruments and iii) inline monitoring of the production of recombinant Lipase B from C*andida antarctica* in *Pichia pastoris* using Glycerol as carbon source.

i) Extended API concentration ranges during granulation and tableting using the PS

The PS is a methodology for preparing calibration sets by adding the changes due to the manufacturing process to NIR spectra of samples prepared at the laboratory, using an algebraic procedure. The PS has successfully included the process contributions during modelling in the central point of API concentration values of diverse formulations, however the properties of this methodology at extreme points of API concentration ranges have not been studied yet. For evaluating such properties, in this work the PS was applied to samples in the range of ± 30% of a nominal API value. Results have shown that the PS performance can be affected by API concentration changes in the studied range, and classical pre-treatments are not enough to overcome this condition.

ii) Comparison of the performance of bench-top and hand-held NIR instruments concerning the identification of NPS

The NPS are 'legal highs' with molecular differences regarding the structures of illicit controlled drugs, whose emergence have expanded the current synthetic drugs market in a very important way. The feasibility of using portable NIRS instruments for the fast identification of NPS have been previously demonstrated, however, their performance has not been faced to the performance of bench-top instruments. Results presented in this thesis expose that, even when models developed using data from NIRS miniaturized instruments are limited in performance regarding those developed using data provided by bench-top instruments, classification models of NPS based on data from hand-held instruments can be useful to make real-time and on-site decisions that can be confirmed later using high performance analytical instrumentation.

iii) Inline monitoring of the production of recombinant Lipase B from *Candida antarctica* in *Pichia pastoris* using glycerol as carbon source.

The use of new constitutive promoters and recycled carbon sources in the recombinant production of industrial proteins, such as lipases, in the cell factory *Pichia pastoris* is advantageous for improving production yields and minimizing the cost of the culture medium. The capabilities of a NIR spectrometer with fiber optic coupling for immersion of a transflectance probe were employed for the inline monitoring of the cultivation mentioned in the headline. Quantitative models have been developed for Biomass, Total protein, Nitrogen and Activity, which have demonstrated better prediction capability during the feed batch stage than during the batch stage. Predictions of glycerol values has been probably affected by the formation of hydrogen bonds in the aqueos medium.

Table of content

OBJECTIVES

The general objective of this thesis is the development of analytical methods for monitoring pharmaceutical and biotechnological processes and for controlling illicit drugs. The analytical technique that has been applied in all the cases is the Near Infrared Spectroscopy, with three different sample presentation modes: off-line, on-site and in-line. The analysis of the data acquired has involved the use of chemometrical methods both for classification and quantification purposes.

Three works have been settled to achieve this general goal:

1. Development of quantitative models based on partial least squares (PLS) regressions for the study of the performance of the process spectrum (PS) at extreme concentration values of active pharmaceutical ingredient (API) in pharmaceutical solid preparations.
2. Development of spectral libraries for identification of new psychoactive substances (NPS) using NIRS data both from hand-held and bench-top instruments
3. Development of quantitative models based on PLS regressions for the prediction of concentrations and activity values of five analytes during the production of a protein (Lipase B) from the constitutive promoter *Candida antarctica* in the yeast *Pichia pastoris* using Glycerol as carbon source.

1. Introduction

1.1 An historical perspective

The establishment of the International System of Units (*Le Système International d'Unités,* SI), during the French Revolution, mid XVII century, opened the door to the valorisation of standardized measurements. This was one of the results of the Industrial Revolution in Europe, which pointed out the age of modern science, especially in chemistry and physics. The SI allowed scientific activities became more precise. Standard reference models of metric units were made, and in consequence, it was required that measuring devices employed for trade and commerce were regularly checked for accuracy against official standard versions. At the same time, efforts for connecting the definitions of the base units to more universally stable properties, were made [1].

This early work on standardization drove important advances in analytical chemistry and its increased impact on society. However, the most relevant progresses were achieved during the Instrumental Revolution, in the two decades between 1955 and 1975, when the analytical profession was transformed by technological developments. During that period, it is possible to find advances in analytical chemistry applied to very diverse needs. The reason for that was the more accurate and precise measurements that became available. According to De Galan, the development and improvement of the instruments for such measurements have been the result of the contributions by four major professional groups, who together have brought the analytical chemistry to its current capability: the inventors, the instrument makers, the analytical chemists and the clients [2].

The inventor, who made the discoveries of natural phenomena, who many times is not aware about its analytical potential at the time of the achievement, starts the chain. The second contributors are the instrument makers, who convert the often rudimentary academic prototype of the novel technique into a reliable device. A team generally completes this task, where the key is the close cooperation between engineers and salesmen. Engineers design the instrument and salesmen collect and transfer insights from their customers before and after purchase. The speed by which the instrumental techniques reached the market varies significantly from one technique to another, and depends on diverse factors [2].

Next, are the analytical chemists, the customers of the instrument makers, who adapt the instruments to their needs. From the beginning of the Industrial Revolution, the main challenges for analytical chemistry have been the reduction of the detection limits and the analysis time, therefore, these have been the focus of the efforts of analytical chemists from that time on. Additionally, the Instrumental Revolution increased the range of available techniques, and expanded the profession from the largely inorganic classical analysis into the domains of the organic chemist and, in recent years, the biochemist. Nowadays it is possible to measure almost any analyte in almost every kind of matrix, fact that has been aided by the push of technology. This great progress has also followed a clear trend: substituting the classical absolute analysis, based on analyte-specific chemical reactions, by instrumental methods, which are all relative. These new methods require a calibration with known standards to convert a physical signal into a chemical concentration. This fact has promoted the rising of new terminology and concepts, particularly in the industrial field, government and statutory regulations. The definition of concepts and procedures devoted to ensuring the traceability of analytical measurements has been prompted to involve the new capabilities of analytical chemistry into the industry [2].

Finally, there is the client, the ultimate receiver of the data provided by analytical laboratories. The clients describe the requirements of new analytical methodologies; they are the ones who outline the conditions under what the analytical chemists must work. Therefore, the feedback obtained from the clients is fundamental for advances in analytical chemistry [2], as the example provided by Near Infrared Spectroscopy illustrates bellow.

1.1.1 The development of Near Infrared Spectroscopy (NIRS)

a) The inventors

Even when it is an undisputed fact that the United Kingdom pioneered the Industrial Revolution in Europe, the fundamental position of the analytical chemistry did not receive enough attention in Britain, regarding other main branches of chemistry, during the XIX century. At that time, academic analytical chemistry was not as widely established in Britain as elsewhere in the world. However, the number of British contributions to analytical chemistry up to the mid–nineteenth century is quite important. Particularly in the field of spectroscopy, the number of contributors is remarkable over this period [3]. These series of contributions started in 1800, with the insights of William Friedrich Wilhelm Herschel, a German musician, teacher and astronomer whom born in

Hannover, on November 15th 1738. W. Herschel became the astronomer of the King George III of the United Kingdom of Great Britain and Ireland in 1782. He settled in Britain after participating in the battle of Hastenbeck, during the War of the Seven Years [4].

Many historians have indicated Herschel as the discoverer of the infrared radiation. The paper entitled "*Experiments on the refrangibility of the invisible rays of the sun*", read in London on April 24th 1800 [5], describes the study completed by W. Herschel on the temperatures of diverse zones of the spectrum. This study allowed the detection of the infrared radiation by means of differences in the temperatures of the visible and invisible zones of the spectrum [4]. Figure 1, is a reproduction of the illustration presented by Hershel in the mentioned paper.

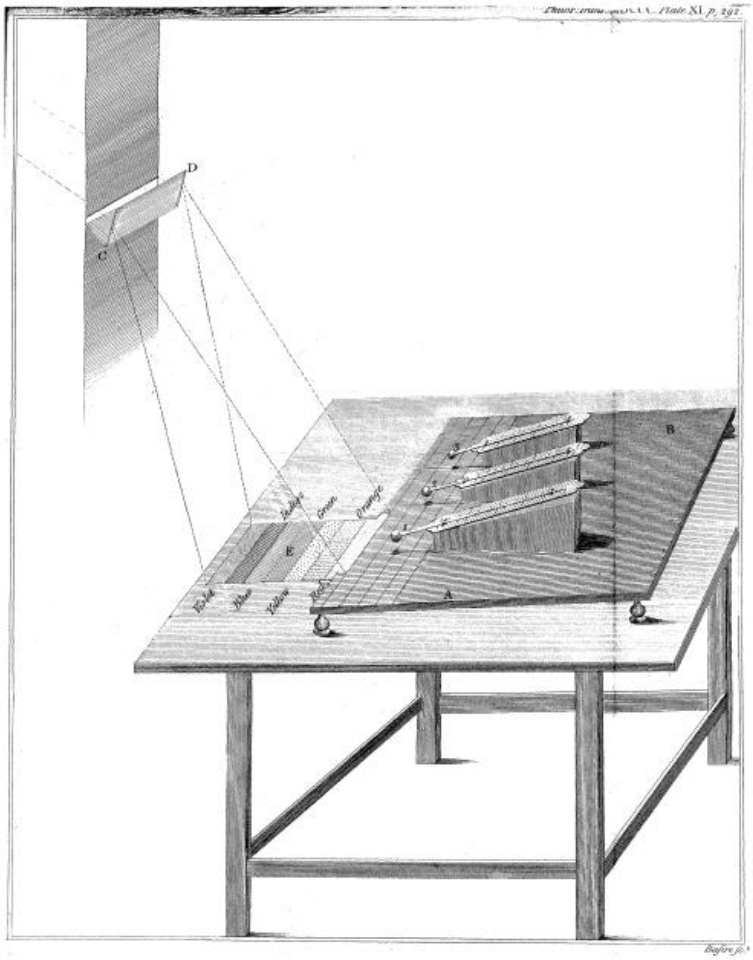


Figure 1. Reproduction of the illustration employed for William Herschel for the exposition of his results on the article *"Experiments on the Refrangibility of the visible Rays of the Sun"*. Taken from reference [5], content downloaded from 158.109.55.24 on Thu, 14 Feb 2019 17:41:49 UTC.

Even though, it is important to keep in mind that the contribution of Herschel was based on the work of the physicist Isaac Newton, who proved over one hundred years before, that white light was made of a mixture of colours that could be separated by a prism. Newton referred to this phenomenon as the spectrum. The efforts of Newton also found antecedents on the theories that earlier scholars of the XIII century stated, which indicated that colours of rainbow were caused by sunlight reflecting through droplets of water [6].

On the other hand, some historians date the discovery of infrared radiation in 1777, based on the concept of radiant heat proposed by Carl Scheele in the famous treatise in which he presented the discovery of the oxygen. Afterwards, Marc-Auguste Pictet reported an experiment with two parallel mirrors, in which invisible radiation was reflected from one mirror to the other, which generated an increment of temperature in the focused mirror. This effect became stronger when the tip of the thermometer was blackened. The work of Pictet was published in Geneva in 1790, but seemingly remained unknown to those who later developed the infrared spectroscopy technique [7].

b) The instrument makers

In any case, after the experiment of Herschel, infrared radiation remained as an object of research instead of a research tool, until the instrument makers became involved in the process. The first on the list of these contributors is Leopoldo Nobili, who reported in 1829 the thermopile, an early infrared detector. Macedonio Meloni, who found that rock salt was much more transparent to infrared radiation than glass, extended the range of the thermopile. This insight allowed Meloni to make better prisms for the detectors. Afterwards, the progress of the infrared techniques slowed in the next years, until the 1840s, when Samuel Langley presented a more precise heat detector named bolometer. The bolometer enabled the measurement of the relatively weak grating spectra and the determination of wavelengths in the infrared ranges by means of a sensitive resistance thermometer connected with a galvanometer. This fact allowed accurate measurement of absolute wavelengths.

Until this point, the research on infrared radiation remained in the domain of physics, because they were the only ones with the required knowledge of the infrared phenomenon and its complex instrumentation. Over the period from 1880 to 1892, pioneer papers on spectroscopy point to the utility of infrared spectra on research on functional groupings in organic molecules. One of these articles was cited by William Coblentz in 1905, when he took up spectral identification of organic compounds and provide the first experimental evidence of the relation between molecular structure and spectral characteristics of substances. His equipment consisted of a rock-salt prism mounted on a mirror spectrometer, with a lamp before the collimator slit and a sensitive radiometer, instead of the bolometer. The contribution of Coblentz provided three main conclusions: i) the configurations between atoms in a molecule are reflected in the spectrum; ii) an increase in molecular weight does no lead to a shift of the absorption maximum; and iii) certain absorption frequencies remain constant for molecular groups even in the presence of other groups in the same molecule. These were the fundamentals for a wide variety of analytical uses of infrared spectroscopy.

However, the experiments of Coblentz did not lead to an increment of chemical analyses on the same lines, probably because the understanding of chemists about the physical bases of spectroscopy at that time were not enough to face this task [7]. By this time, investigations of molecular structure by means of infrared spectra began to appear slowly, while theoretical physics experienced the emergence of conceptual issues on the interaction between matter and radiation. In the 1920s, the development of quantum mechanics, prompted by Max Planck, opened the door to the determination of the vibrational frequencies. Based on the work of Planck, the vibrational energy levels of molecules were developed. A limited group of symmetrical molecules of low molecular weight outlined the starting scenarios. The arrival of Raman spectroscopy supported the refining of the process, enabling the use of the two techniques together in the study and interpretation of the energetics of simple molecules in terms of their rotational and vibrational behaviour. Unfortunately, these preliminary efforts were not enough to understand the vibrational bands of heavy, complex and asymmetric molecules [7].

c) The analytical chemists and clients

The challenge of complex asymmetric molecules was faced by a collaborative group created between the US National Bureau of Standards (NBS) -where Coblentz was an active researcher- and graduated students of the Johns Hopkins University in Baltimore. Scientists of NBS were interested in infrared radiation for complementing the identification of hydrocarbons in their systematic study of the chemical composition of petroleum. The Johns Hopkins researchers had the objective of understanding the infrared radiation in terms of molecular structure and theoretical chemistry. Findings of this collaborative group prompted the American Petroleum Institute (API) to apply the technique systematically in a study of chemical composition of petroleum. This was the base for the work published by F. Rose in 1938, which brought two new important ideas: iv) different structural groups of hydrocarbons have absorption maxima at different frequencies; and v) a given grouping has a constant absorption intensity at each of its characteristic frequencies [8]. From that time, the infrared spectroscopy became a fast and reliable tool for monitoring complex organic reactions.

However, this important impulse was mainly dedicated to the region of the mid infrared, MIR (2500-16,000 nm), due to the higher amplitude of the fundamental vibrations collected in this region, compared to the low amplitude of vibrations acquired in the near infrared region, NIR (700-2500 nm). Indeed, when the World War II started, mid infrared spectroscopy had a firmly domain of the major chemical and petroleum companies. In the universities, few groups of physicists dedicated to theoretical chemistry used the technique, but organic chemists were not interested in completing the delicate and labour-intensive calibrations and adjustments that the technique still required. Because of that, the early evolution of the technique took place at the industry, which could afford the time and money required to use it. At this point, the most experienced company in production of optical devices for the industry, Perkin Elmer, became the most relevant instrument maker for the technique. Perkin Elmer made the first compact and mass-produced infrared instrument. This fact contributed to the standardization of mid infrared spectroscopy and its diffusion in petroleum refining.

Because these early instruments manufactured by Perkin Elmer were focused on combining the mid infrared region and the visible region, the near infrared remained ignored until the mid-1950's, when Wilbur Kaye, presented three articles, using Beckman Instruments. The effort of Kaye served as a base for the description of NIRS as a technique able of provide relevant structural information [9]. Furthermore, in 1954, the Applied Physics Corporation (Monrovia, California), presented a double bean spectrometer designed to operate over the wide spectral range of ultraviolet, visible and near infrared wavelengths (185-870 nm), the Cary 14. The Cary 14 was the successor of the Cary 11, the first commercially available ultraviolet visible spectrophotometer [9].

However, at that time, the motivation for the construction of the instruments remained far from considering the whole vibrational spectral region. Therefore, the emergence of NIRS into the analytical world was delayed until the 1960's when Karl Norris, a researcher of the U.S. Department of Agriculture, built a low-cost monochrometer using a piece-interference filter, useful to vary the wavelength in a simple way [10]. In this sense, and according to the structure proposed by De Galan, besides the earlier efforts of Coblentz, Karl Norris was the first analytical chemist implicated in the development of the NIRS, and the Department of Agriculture of the USA the first client who provide feedback for the enhancement of the technique.

The initial efforts of Norris were dedicated to the identification of wheat flour and the measurement of moisture using transmission as acquisition mode. Since the measurement in transmission mode required the use of carbon tetrachloride (for the sample to become transparent and allow the measurement with the still very raw instrument) the application of the technique into the food industry was not possible due to the health hazard of this solvent. This was the reason why Norris made internal changes in the instrument to measure diffuse reflected radiation instead of transmitted radiation using the same wavelengths. Afterwards, new improvements were added by the coupling of the NIR spectrometer to a computer where a software was developed to collect and analyse NIRS transmission and diffuse reflection data in the range 400 – 2600 nm [10].

Because of differences in the particle size of whole wheat, Norris and its collaborators noticed differences in their initial spectra and realized that these differences were due to surface reflectance. To overcome this difficulty, they developed an interactance probe for separating the source fibres from the collecting fibres by means of a thin metal strip. This improvement allowed the measurement of fat content through the human skin [11]. The NIR spectrometer that Norris et al. assembled was unique; therefore, it attracted scientists from all over. Figure 2, shows a picture this instrument.

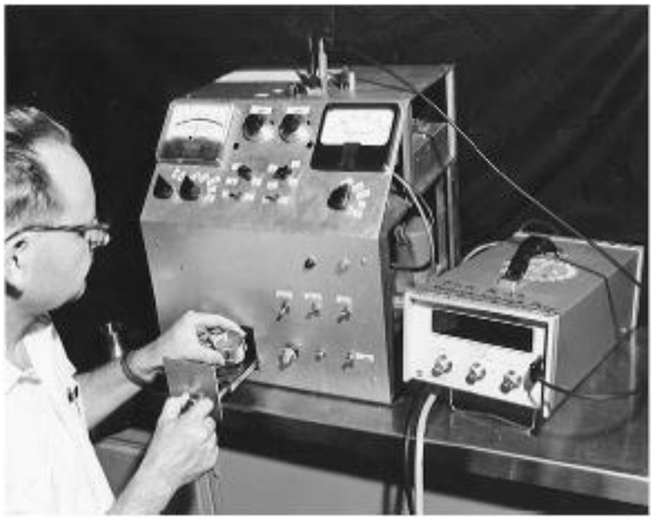


Figure 2. Reproduction of picture of the first NIR instrument developed by Norris et al. for moisture measurement of wheat flour. Taken from reference [10] (doi:10.1016/j.trac.2010.01.003).

These scientists later contributed in spreading the knowledge of NIRS and developing new applications around the world. Canada was the first country to guarantee the protein content of wheat using NIRS in substitution of the Kjeldahl method in the 1970s. In September 1973, the first results of this “real world” application of NIRS were presented at the American Association of Cereal Chemists conference in St. Louis, MO, USA. The same year, an instrument named “Automated Digital Analyzer”, was the first NIRS instrument that allowed the adjustment of optimum wavelengths ranges. Phill Williams was at that time Chemist-in-charge of Protein-testing of the Canadian Grain Research Laboratory (GRL) Kjeldahl laboratory, and supported the transition to the new instrument. The “Automated Digital Analyzer” was digital computer tracked, and it worked without a break, 24 hours-a-day, 7 days-a- week until 1992. Over this period, this instrument completed about 11 million tests for protein and moisture, without once breaking down [10]. This fact represented remarkable savings in laboratory spences, as well as an important reduction of Kjendals reactives wastes.

Afterwards, diverse companies as Neotec, Technicon, Tecator (now Foss Analytics) expanded and enhanced the NIRS instruments in such a way that for 2010, around a 90% of the wheat world-wide was sold on the basis of protein testing by whole-grain NIR spectrometers. Additionally, over this period, the applications of NIRS to pharmaceutical requirements began to be explored with successful and increasingly more expansive results.

d) Evolution of the presentation of the sample to the instrument

Most of the current industrial applications of NIRS have been developed based on off-line measurements on raw materials and finished products for quality control purposes. However, it is worth mentioning that technique traced by itself the “Five eras” described by *Callis* et. al. for the evolution of the Process Analytical Chemistry (PAC): off-line, at-line, on-line, in-line and non-invasive eras [12]. The PAC involves many diverse analytical techniques, but NIRS is probably one of the most representative of this concept. In fact, currently, the five eras of its development are described as diverse possibilities of instrumental configuration for NIRS, because depending on the application, one or more of these ways of presentation of the sample to the instrument can be useful for solving customised analytical needs.

The **off-line** configuration requires a starting sampling procedure for collection of a set of samples to be analysed in a laboratory facility, in a location different to the process plant. **At-line** configurations are related to the analysis inside of the process plant, because the instrument is installed there; however, the sample is analysed out of the process line. Therefore, the sampling procedure is still required. **On-line** measurements are those completed over samples taken from a process line, analysed by means of loops or lines specially created for the analysis with NIRS, which later return the sample to the process. **In-line** measurements are achieved using probes specially developed for immersion in process reactors or manufacturing lines. Finally, **non-invasive** measurements are completed using accessories designed for applications were the instrument acquires the spectra by means of a transparent window or directly over the product, depending on the process conditions. This last kind of configuration is particularly useful for solid samples and is the most representative of PAC applications of NIRS.

The PAC, nowadays also named process analytical technology (PAT), pretends to provide qualitative and quantitative information about a chemical process, with the aim of optimizing the use of energy, time and raw materials [12]. The early origins of this idea was a specialized form of at-line, real-time analysis in 1937, according to a British Intelligence Operations Subcommittee Report, which exposed that Germans used specialized instrumentation for process control in their chemical industry [13]. PAT is valuable because allows the availability of updated information for the opportune decision making about process variables, that the natural delays of off-line measurements makes hard. Informed and opportune process changes are key for optimizing safety, quality, and production cost efficiency, aspects of relevant impact for all the industries.

Based on this fact, recent developments on NIRS instrumentation have been devoted to the design and manufacturing of probes and accessories for in-line and non-invasive process measurements, as well as the reduction of the size of the instruments. Such progresses have been supported on advances on interferometers and dispersive monochromators, the diode-array detectors and the new perspective provided by the Raman spectroscopy in the 1990s.

Unfortunately, the most important milestone in the reduction of the size of the NIRS instruments was a tragic situation. Portable and handheld optical spectrometers, started to have a relevant presence in the market with the destruction of the World Trade Center (New York City), on September 11th, 2001. The urgent demands for screening, detection and identification of explosives, hazardous materials among other chemical substances, compelled the fast development of these instruments[14].

Currently available small NIRS spectrometers can be roughly classified intro three groups:

- Small versions of laboratory spectrometers: transportable instruments oriented to faster data acquisition.
- Process analyzers: for generation and transmission of qualitative or quantitative information to a process controller.
- Dedicated field analyzers (handheld spectrometers): developed for providing answers to non-specialists.

The recent availability of low-cost sensors and other electronic components has led to the development of low-cost portable spectrometers [14].

1.2 Physical fundamentals

The region of the electromagnetic spectrum limited by the visible and the microwave regions (700-111,000 nm) is named the infrared (IR) region. This spectrum is the result of the absorption of light by matter and is associated to stretching and bending modes that mainly occur in intramolecular covalent bonds of organic molecules, although hydrogen bonding and intermolecular interactions can affect it as well.

The IR region is divided in three sub-regions:

- The far infrared (FIR): Located between 16,000 and 111,000 nm. Primarily employed for rotational spectroscopy, with wide applications in astrophysics.
- The mid infrared (MIR): Located between 2,500 and 16,000 nm. Traditionally employed for molecular characterization due to the chemical information involved in the sharp of the spectral profiles produced.
- The near infrared (NIR): Located between 700 and 2500 nm. Associated to overtones and combination bands of the fundamental vibrations of the MIR region.

When the MIR radiation interacts with matter, diverse molecular stretching and bending motions are induced in the molecules. Such movements depend on the frequencies of the radiation. After these fundamental vibrations take place, further overtones and combination bands occur as physical consequence of these starting movements [15]. These overtones and combination bands constitute the NIR spectrum.

To understand this phenomenon, it is useful take into account the properties both of the electromagnetic radiation and the vibrations in molecules.

1.2.1 Properties of electromagnetic radiation

The infrared radiation may be considered as a simple harmonic wave. Therefore, this radiation can be defined in terms of the properties of the sine wave and the distance travelled in a complete cycle of such *sine* wave. Any simple wave of radiation undulates interconnecting electric and magnetic fields, which interact with matter to generate a spectrum. As any simple harmonic motion, the properties of the wave of radiation can be can be defined by:

$$y = A \sin \theta \qquad \textit{Equation 1}$$

where y is the displacement with a maximum value A, and θ is an angle that varies in the range from 0 to 2π radians. Therefore, the travelling wave follows a circular path of radius A, describing an angle $\theta = \omega t$ radians, t seconds after passing the maximum point of its vertical displacement:

$$y = A \sin \omega t \qquad \textit{Equation 2}$$

After 2π/ω seconds, the cycle is completed. That means that in one second the undulatory pattern is repeated ω/2π times. The expression of the wave regarding the time is known as frequency (v)

of the wave. Considering the frequency, the basic equation of the wave movement can be rewritten as:

$$y = A \sin 2\pi \nu t \qquad \textit{Equation 3}$$

The undulation also generates horizontal displacement while completing each cycle, which is known as the wavelength (λ). Expressing the equation 3 in terms of such displacement, in distance instead of time, involves the substitution of $t=l/\boldsymbol{c}$, where l is the distance covered by the wave in time t at velocity $\boldsymbol{c}$ ($\boldsymbol{c}$ is the universal constant for the velocity of light *in vacuo*). Based on that, the wavelength can be defined as:

$$\lambda = \boldsymbol{c}/\nu \qquad \textit{Equation 4}$$

The infrared radiation may also be described is in terms of wavenumber ($\bar{\nu}$). The wavenumber is defined as the reciprocal of the wavelength expressed in centimetres:

$$\bar{\nu} = \frac{1}{\lambda} cm^{-1} \qquad \textit{Equation 5}$$

The $\bar{\nu}$ is commonly considered as the number of waves or cycles per centimetre of radiation. Conventionally, spectroscopists describe the position of an infrared absorption band in terms of its wavenumber, since it is directly proportional to frequency (ν = c$\bar{\nu}$) and this unit is more easily related to the energy changes associated to transitions between different vibrational states [16].

1.2.2 Properties of vibrating molecules

Even when the wave model was valuable to understand the properties of radiation, it is no longer useful to account for the phenomena associated to the absorption or emission of energy in the near infrared region. For these processes, it is required to visualize the electromagnetic radiation as a stream of discrete particles (photons) with an energy proportional to the frequency of the radiation. Based on this view, the NIRS can be considered a consequence of both electronic and vibrational transitions.

In order to understand these transitions, it is important firstly to state that when molecules absorb the infrared radiation, this radiation generates vibrations in their individual bonds. Such vibrations can be generated in both intra and intermolecular bonds. Each molecular bond can be individually considered as a diatomic oscillator, which from the perspective of the harmonic oscillator has a potential energy V that can be defined as:

$$V = \frac{1}{2}k(r - r_e) = \ \frac{1}{2}kx^2 \qquad \textit{Equation 6}$$

where k is the force constant of the bond, r is the internuclear distance, r_e is the internuclear distance at the equilibrium state and x = $(r$ - $r_e)$ is the displacement coordinate. The harmonic oscillator model describes the vibrational frequency v such that the curve for the potential energy has a parabolic shape, symmetrical about the equilibrium bond length, $r_{e.}$ The definition of v is described in equation 7:

$$v = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \textit{Equation 7}$$

where μ is the reduced molecular mass, such that $\mu = \ m_1m_2/(m_1 + m_2)$, and m_1 and m_2 are the masses of the nuclei involved in the bond [17]. However, most of the electronic transitions exposed in the NIR region are due to d-d transitions, charge-transfer transitions, and π–π transitions of conjugated systems of C-H bonds, which are forbidden transitions within the harmonic oscillator case. Experimental observations give evidence that molecules are not ideal oscillators. The first evidence is that their vibrational energy levels are not equally spaced, as can be seen in Figure 3. Therefore, the hot bands do not have exactly the same frequency as the fundamental band. The second evidence is that overtone transitions, like v = 0 to v = 2, 3, 4 and so on, are allowed [18].

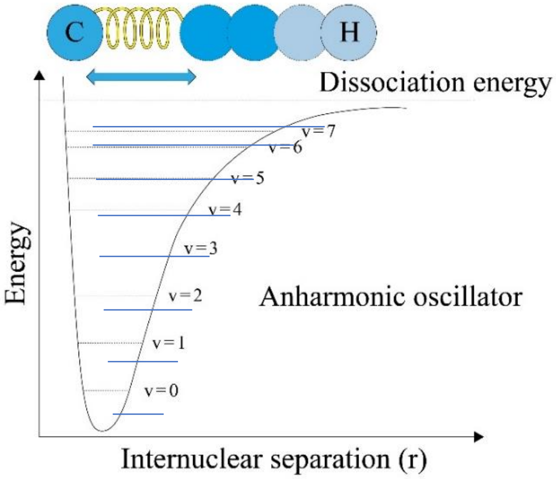


Figure 3. Illustration of the energy diagram of vibrational modes calculated as an anharmonic oscillator. Adapted from reference [15].

Because of that, it is necessary to introduce the term anharmonicity. The anharmonic behaviour is key for understanding NIR spectrum in terms of both intensity and frequency and it can be expressed by means of two effects. One of these effects is called the *mechanical anharmonicity*, which is due to the cubic and higher terms of displacement coordinates in the potential-energy expression presented in equation 8:

$$V = \frac{1}{2}kx^2 + k'x^3 + \cdots k' \ll k \qquad \textit{Equation 8}$$

Equation 7, is employed in the Schrödinger equation to deduce the energy levels of the allowed states of the anharmonic oscillator. The solution has been obtained based on an approximation that can be written as:

$$G(v) = E_{vib}/hc = \bar{v}\left(v + \frac{1}{2}\right) - x_e\bar{v}\left(v + \frac{1}{2}\right)^2$$

$$= \bar{v}\left(v + \frac{1}{2}\right) - X\left(v + \frac{1}{2}\right)^2 \qquad \textit{Equation 9}$$

where *h* is the Plank constant, x_e is the anharmonicity constant and $X = x_e$. As a consequence of equation 8, the energy levels are not equally spaced, as is showed in Figure 1. The other effect useful for the expression of the anharmonic oscillator behaviour is named the *electrical anharmonicity*, which is responsible for the appearance of overtones corresponding to transitions between energy levels that differ in two or three vibrational quantum number units in the infrared spectra. The electrical anharmonicity is due to the effect of square and higher terms in the dipole-moment expression:

$$\varepsilon = \varepsilon_0 + \left(\frac{d\varepsilon}{dx}\right)_e x + \left(\frac{d^2\varepsilon}{dx^2}\right)_e x^2 \quad + \cdots \qquad \textit{Equation 10}$$

where ε is the energy of each energy level and ε_0 the energy at the fundamental level. Based on equation 9, it is possible to see that, for the anharmonic oscillator, the frequencies of the overtone absorptions are not exactly 2, 3, … times the fundamental absorptions [17].

Some relevant consequences of the anharmonic behaviour of the near infrared absorption are:

- ✓ The intensity of the bands in the near infrared region is much weaker than the fundamental absorption bands in the mid infrared region.

- ✓ Due to a number of overtones, combination bands and Fermi resonances overlap each other in the NIR region; the interpretation of the spectra is not straightforward.
- ✓ Bands associated to functional groups with hydrogen atoms dominate the NIR spectra, since the anharmonic constant of an X-H bond is usually high. On the other hand, C=C vibration does not generate bands in the NIR region. The information from these bonds is obtained by means of combination modes of C-H groups linked with such C=C bond.
- ✓ The shift and intensity changes due to hydrogen bonds and interaction between molecules is more important in NIR bands than in MIR bands.
- ✓ For bands associated to the first overtones of the X-H bonds (X = O, N), streaching modes of monomeric species exhibit higher intensity regarding the corresponding bands of polymeric species [19].

The anharmonic behaviour also makes that the NIR spectroscopy currently holds significant advantages over MIR and Raman spectroscopy for physical chemistry and molecular science investigations. Studies of NIR anharmonicity and vibrational potentials are nowadays fundamental for understanding molecular structures, and combination of NIR data and quantum chemical calculations support developments on anharmonicity and vibrational potentials. This is probably because studying MIR and Raman spectra involves only the description of the dipole moment and the polarizability respectively. However, describing the NIR spectra in terms of both intensity and frequency comprises the anharmonicity of vibrational potential and nonlinearity in the dipole moment function [19].

1.2.3 Overtones, combination bands and Fermi resonance

As has been explained above, overtones and combination modes are the bands exposed in the NIR spectrum, therefore, it is important to describe them. The overtones are forbidden transitions in harmonic oscillator approximations, whose intensities decrease in an exponential manner with the increase in the vibrational quantum number [19]. A valuable analogy for understanding NIR overtones is the ringing of a bell, illustrated in Figure 4. When the bell collides with another object, the first sounds are loud and highly clear in audible frequencies. These are the fundamental notes of the bell sound, with a high amplitude. As the bell is left to vibrate, the sound intensity decreases rapidly, and the sound becomes each time gentler. Oscillating sounds only can be heard after the fundamentals have declined. These further vibrations are known as the overtones of the fundamental frequencies, and typically occur at integer values of the fundamental vibration. That is, if the fundamental vibration occurs at a frequency value *f*, theoretically, the first overtone occurs at 2*f*, the second overtone at 3*f*, and so on.

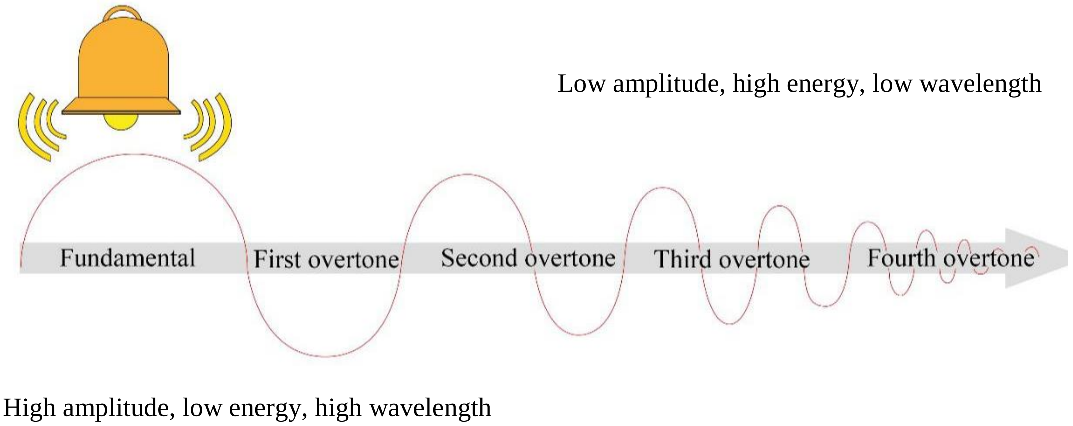

Low amplitude, high energy, low wavelength
Fundamental
First overtone
Second overtone
Third overtone
Fourth overtone
High amplitude, low energy, high wavelength

Figure 4. Drawing of changes in the amplitude between fundamental vibration and further overtones in the ringing of a bell. Illustration for analogy with overtones observed in the NIR region. Adapted from reference [15].

When a molecular bond absorbs energy at a fundamental frequency *f* in the MIR region, then at approximately 2*f*, the first overtone band of the fundamental frequency will occur in the NIR region. As with the case of the bell ringing, the overtone frequency has an intensity of approximately an order of magnitude less than the fundamental. The following overtones will be each time of smaller amplitude by an order of magnitude from the last overtone. Generally, these overtone vibrations have low molar absorptivity coefficient regarding the fundamental frequency [15]. This is the reason why no sample preparation is required using NIRS, because the intensity of the signal is so small that samples do not need to be diluted as is required in the MIR region.

The molar absorptivity coefficient is a ratio that establishes the absorbance at a particular molar concentration of an analyte, using a known pathlength at each wavelength. The correlation between these three terms is obtained from the equation 10, which is based on the law of Beer:

$$\varepsilon = \frac{A}{cl} \qquad \textit{Equation 11}$$

where ε is the molar absorptivity coefficient, *A* is the absorbance, *c* is the concentration of the analyte, and *l* is the pathlength. According to the SI, the values for the pathlength should be presented in meters, but it is a current general practice to describe it in centimetres. The measurement conditions of the analyte, such as solvent, pH and temperature affects the molar absorption coefficient [20]. Therefore, when possible, it is a good practice to use single standard for determining the molar absorptivity coefficient, instead of assuming the adherence to the law of Beer-Lambert, or using literature values [21].

The stretching frequencies have a significant contribution to the overtone frequencies in NIR. Even though, strong bending vibrations also contribute at this region. Combinations occur when the overtones generated in the MIR combine to form bands of higher intensity than would occur from the overtone alone, i.e. combinations may be considered as the average frequency of two adjacent molecular vibrations in a molecule. In this case, the proximity between atoms is more important than the nearness of the energy levels [22].

Finally, it is important to highlight in this section that, due to the large number of vibrational energy levels in polyatomic molecules (3N-6 energy levels), often two of these levels have practically the same energies. This proximity allows the Fermi resonance may occur. When the Fermi resonance takes place, the energy levels are subject to a repulsion, moving up one of the levels of energy and the other one down, by the same magnitude. The magnitude of the repulsion (i.e. the magnitude of the Fermi resonance) is directly proportional to the anharmonicity related to the interacting levels. It is difficult to observe the Fermi resonance in the NIR region, because it used to be overlapped under the broad overtones and combination bands, however, it is worth mentioning that it is a kind of effect that contributes to the bands in the NIR spectrum [22].

Based on the overtones and combination bands, the region of the NIR spectrum can be divided as following:

1. The first overtone and combination band region: Located between 2000 and 2500 nm. In this region, combinations occur between the first overtones of the fundamental bands (at a frequency 2*f*) and typically have the higher molar absorptivity.
2. The second overtone and combination bands region: Located between 1100 and 2000 nm. Involves combination bands strong enough in intensity to generate bands in this region.

3. The third and fourth overtones region: Located between 700 and 1100 nm. Usually the molar absorptivity of these bands is too small to reveal any combination bands of practical use [17].

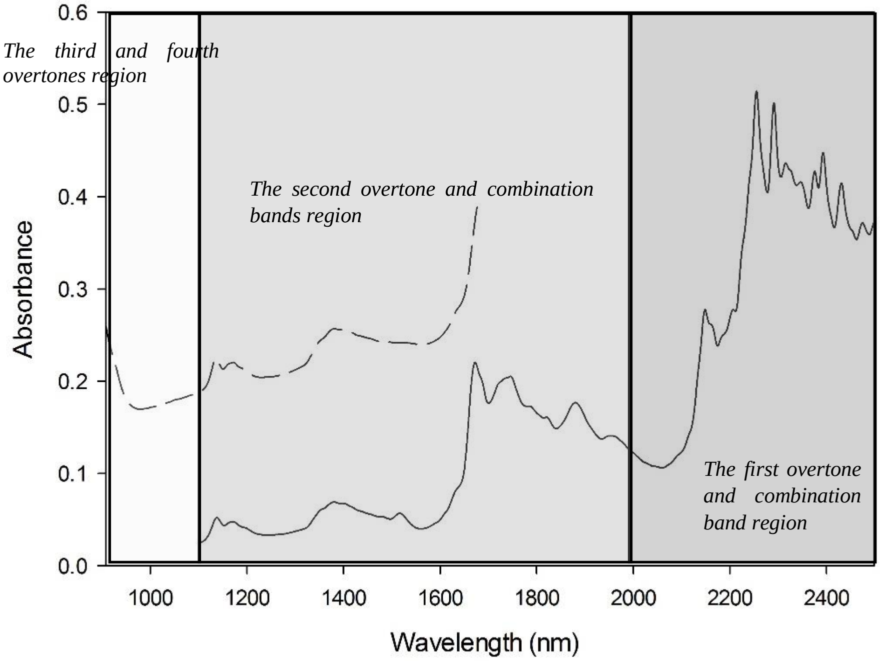


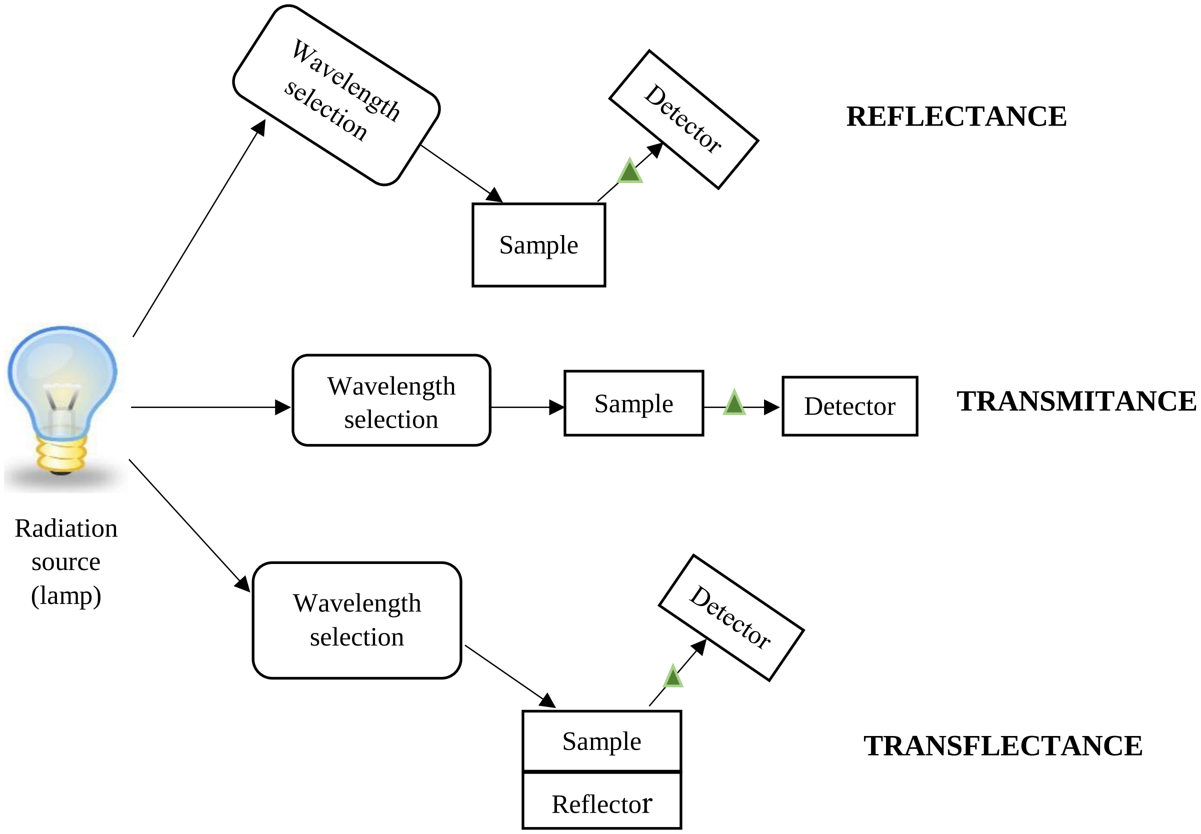
Figure 5. Illustration of regions in the NIR spectrum. Spectra of solid samples acquired in reflection mode with two instruments of different kind of detectors.

The spectral range available of a particular NIR spectrum depends on the kind of detector used for its acquisition. Figure 5, illustrates the above mentioned regions of the NIR spectrum. As can be seen, in this graph, two spectra have been required to expose the three regions, due to not all the instruments have detectors able to acquire data from the whole NIR region. Therefore, in this case, two different instruments have been used.

Sections 1.3 and 1.4, presents the diverse NIR instruments and their capabilities, among other insights on the acquisition of the NIR spectra.

1.3 Acquisition modes

In the NIR spectral range, the interaction of the radiation with the sample takes places in three different ways: reflectance, transmittance and transflectance. These are the three acquisition modes available for the technique. In general, the physical state of the sample orientates the selection of the acquisition mode, however, depending on the whole analysis conditions, feasibility studies can be required for the definition of the proper acquisition mode. The main difference between the three acquisition modes is the position of the detector regarding the sample. Figure 4, shows the general disposition of the parts of a NIR instrument for the different acquisition modes.



▲ Wavelength selection devices can be also included at these points

Figure 6. General description of instrumental configurations for the different acquisition modes in NIRS: reflectance, transmittance and transflectance.

In all the cases, the analytical signal obtained is a logarithmic function of the apparent absorbance, described in equation 12:

$$A_i = log \frac{1}{X_i} \qquad \text{Equation } \mathit{12}$$

where A_i is the absorbance calculated for the ith wavelength of the spectrum, and X_i is the reflectance (R) or transmittance (T) of a sample at the ith wavelength. The absorbance is expressed on a unitless scale and is represented by the symbol AU (absorbance units). Since the scale is logarithmic, each AU represents an order of magnitude less light intensity than the incident light source, as is illustrated in Figure 7. This means that each absorbance value represents 10 times less light incident on the detector regarding the incident radiation, and an absorbance value of 5 represents 100,000 times less light detected regarding the incident radiation.

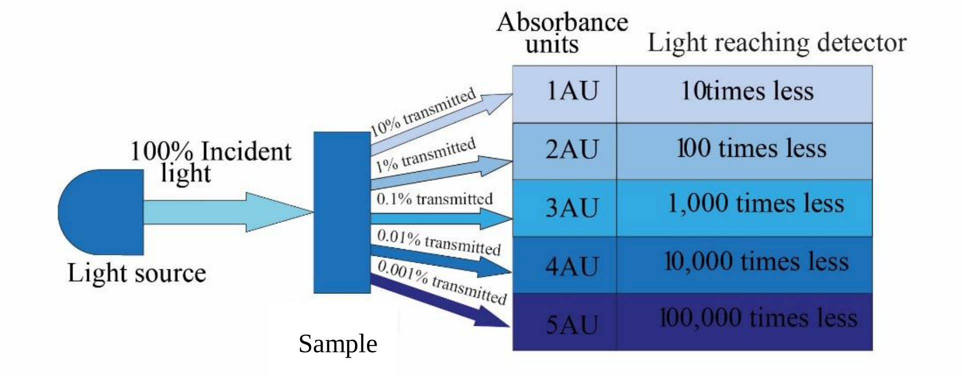


Figure 7. Diagram for illustrating the effect of the logarithmic function of the absorbance. Adapted from reference [15].

A general flow for acquiring a spectrum using an NIR instrument typically comprises:

1. Collection of a dark current (DC) spectrum. This collection is usually completed where the light source of the instrument is turned off; therefore, the electronic noise of the detector is measured for all the wavelengths.
2. Collection of a reference spectrum. For transmission mode, this spectrum is the signal from the light source without interaction with any sample. For reflectance mode, the

reference spectrum, the reference spectra is the result of the reflection of the radiation from a standard highly reflective material measured for all wavelengths.

3. Collection of the sample (Sam) spectrum. This action is completed by placing the sample in front of the light source and collection the radiation either transmitted through or reflected off the sample.

The resulting raw transmittance or reflectance scan is obtained using the following ratio [15]:

$$X_i = \frac{Sam_i - DC_i}{Ref_i - DC_i} \qquad \text{Equation 13}$$

In the case of diffuse reflectance, the apparent absorbance is expressed as Kubelka-Munk units, as will be explained bellow.

1.3.1 Diffuse reflectance measurements

When the radiation interacts with a solid sample, it can be reflected by specular or diffuse reflection. The specular reflectance has been described by Fresnel and is directly proportional to the value of the absorption coefficient at a particular incident wavelength. The specular reflectance takes place in only one direction (the incidence plane), and it predominates when the penetration of the radiation is too small regarding the wavelength, or when the dimensions of the reflectance surface are larger than the wavelength incident [23]. On the other hand, the diffuse reflectance is a consequence both of absorbance and scattering processes, it takes place in all the directions and predominates when the materials of the reflecting surface are weakly absorbent at the incident wavelength. The diffuse reflectance is also produced when the penetration of the radiation is larger regarding the wavelength [22].

In general, reflectance measurements involve components from both the specular and the diffuse reflection. In NIRS, the components of the specular reflection provide very few information about the composition of the sample, then its contribution can be minimized with the position of the detector regarding the sample. Conversely, the diffuse reflectance is the responsible of most of the useful information acquired using NIRS, therefore this phenomenon is the base of measurements using reflectance mode. The diffuse reflectance has been explained by the theory initially exposed by Kubelka and Munk in 1931. This theory is based on several assumptions, the most relevant are that the incident radiation in a scattering medium undergoes simultaneously absorption and scattering processes. Based on that, the reflected radiation can be described in terms of the absorption constant k and the scattering constant s. In the case of matte samples with an infinite thickness, the Kubelka-Munk function can be expressed as [22]:

$$f(R_\infty) = \frac{(1-R_\infty)^2}{2R_\infty} = \frac{k}{s} \qquad \textit{Equation 14}$$

where R_∞ is the absolute reflectance of the sample, which is the fraction of incident radiation reflected. In practice, the relative reflectance R is used instead of R_∞. The R can be defined as the ratio between the intensities of radiation reflected by the sample and a standard material. The standard used to be a highly stable material, with a large and relatively stable constant of absolute reflectance in the NIR region. Some common examples are the Teflon, the Barium sulphate, Magnesium oxide, and alumina ceramic plates of high purity.

Due to the term R can be related to the concentration of the analyte, the equation 14 can generate a graph with a slope that include the *1* as a possible solution. Even though, in cases where the matrix absorbs or when the absorption bands of the analytes is too intense, the diffuse reflection of the samples does not comply the Kubelka-Munk equation and the graph of f(R) vs concentration is no longer lineal. Because of that, it is accepted that the equation of Kubelka-Munk, as the law of Beer-Lambert, is a limit equation, which can only be applied to absorbing bands of low intensity. This is the case of the absorbing bands in NIRS, however, due to it is not possible to isolate the absorption of the analyte from the absorption of the matrix –which often absorbs with high intensity at the same wavelength of the analyte- deviations of the equation 14 occur [22]. To overcome this condition, diverse studies have been advanced, due to the growth in diffuse reflectance applications have promoted the progress in the understating of the physics behind the phenomenon [24].

From a practical point of view, an alternative commonly used is the application of the following relation between the concentration and the relative reflectance equivalent to the law of Beer-Lambert:

$$log\frac{R_{standard}}{R_{sample}} = log\frac{1}{R_{sample}} + logR_{standard} \approx \frac{ac}{s} \qquad \textit{Equation 15}$$

where a is the molar absorptivity of the sample and c is the concentration. For monochromatic radiation, the log $R_{standard}$ can be considered constant, and the equation can be written as:

$$A = \log \frac{1}{R} = a'c \qquad \textit{Equation 16}$$

where R is the relative reflectance, and a' is a constant of proportionality. Even when this expression does not involve the theoretical bases of Kubelka-Munk advances, it provides useful results in conditions often employed in diffuse reflectance.

1.3.2 Transmitance and transflectance measurements

Similar to the case of diffuse reflectance measurements, in the NIR region this equation can be limited by the effects of hydrogen bonds, molecular complexity among other processes. This is the reason why the analysis of solid samples using transmission must consider that the radiation can undergo diffuse reflectance and in this case log 1/T is no longer representative of the attenuation of the radiation by absorption. From a practical point of view, the analysis of solid samples is mainly completed using diffuse reflectance and transmission and transflectance acquisition modes are employed for liquid and semisolid samples. The transflectance mode is a variation of the of the transmission mode. In this case the transmittance is recorded after passing through the sample twice (the radiation travels a twice as far pathlength). The second trail is completed thanks to a reflector located behind the sample, which generates a second travel of the radiation through the sample before reaching the detector.

1.4 Instrumentation

The basic internal components of NIR spectrometers are the radiation source, a system of wavelength selection, a sample chamber and the detector. Differences in the materials and characteristics of each of these parts define the final capabilities of each instrument. The following paragraphs details features of these components.

1.4.1 Radiation source

In general, there are two kinds of radiation sources employed in NIRS: sources of whole range and sources of reduced range. For sources of whole range, the model most commonly used is the halogen lamp with Tungsten filament and quartz window. Halogen lamps provide high intensity radiation and covers the electromagnetic spectrum of NIRS in a wide range, from 320 to 2500 nm.

For sources of reduced range, the Light Emission Diodes (LED) lamps are the most commonly employed [25], [26]. A LED lamp produces light using one or more LEDs. The LEDs are semiconductors that recombine their electrons with electron holes while the current flows through it. The radiation is emitted thanks to the energy released in the form of photons (photoluminescence) [27]. The semiconductors more used in NIRS are the GaAs, which emit in the range ~ 645-830 nm, and the InGaAs, which provide radiation until ~ 1700 nm. Despite the region under 1700 nm exposes only the absorption of the third and fourth overtones, and is limited to C-H, O-H and N-H groups, portable and hand-held instruments take advantages of NIR in this region for diverse applications. The diverse designs and arrangements of sources of radiation of portable NIRS instruments currently available have been carefully reviewed by Crocombe in [14].

1.4.2 System of wavelength selection

Except instruments based on LEDs devices as radiation source, NIR spectrometers require systems for selecting a bandwidth smaller than the whole region provided by the lamp. Based on the fundamentals of their optics, such systems can be classified as dispersive and nondispersive. The monochromators are the dispersive systems more widely used in NIRS. In general, monochromators involve a light inlet, two collimators, a dispersive element and a light outlet. The light inlet allows the radiation reaches the first collimator as a tight beam of light. The first collimator makes the beams of radiation parallel which each other before the dispersive element. Then, the beams are focussed to the light outlet thanks to the second collimator [21]. The monochromators are characterized by the dispersive element, which can be a prism or a diffraction grating. The diffraction grating is the dispersive element more employed in NIR spectrometers [28], probably because of its clear advantage of varying from 3 to perhaps 100 the dispersion of the prism materials [29]. The grating diffracts each wavelength of the incident polychromatic radiation to diverse angles by means of the streaks of this surface, generating both constructive and destructive interferences. The discrete beams that the monochromator produces are transformed into a wider range of wavelengths using an engine, which ensures that each time one wavelength is focused on the light outlet, sweeping in this way the whole NIR spectral range.

The set of nondispersive systems of wavelength selection available for NIRS is more diverse. There are optic filters, Acusto-Optic Tunable Filters (AOTF), and interferometers. Optic filters are semi-transparent elements located between the polychromatic source of radiation and the sample, which allow the pass of only certain wavelengths. The simplest optic filters are the absorption filters, wherein the selection of the wavelength depends on the constitutive material of the filter

[22]. Another kind of filters are the interference filters, also named Fabry-Perot, which are based on the optic interference. In these filters, the wavelengths transmitted depend on the refraction index of the material and the thickness of the filter. Fabry-Perot filters transmit more radiation than absorption filters, with narrower bandwidths [30].

The AOTF are based on the interaction of the radiation with sound waves and provide the capability of tunning the spectral bandpass electronically. An AOTF has no moving parts and is basically a crystal (commonly of TeO_2) sensible to acoustic waves, at radiofrequencies (RF), which are used to separate a single wavelength from a broadband or polychromatic light. The selection of the wavelength depends on the frequency of the RF applied to the crystal and is independent of the geometry of the device. Two piezoelectric transductors are located at each side of the crystal for transforming the RF into an acoustic signal. When the polychromatic radiation impacts the crystal, two beams of monochromatic light polirazed (light waves with vibrations in a single plane) are generated, but only one is used with analytical purposes. The wavelength of both beams is the same and depends on the speed and frequency of the sound wave, the dimensions and the birefringence (double refraction capability) of the crystal. AOTF are fast, provide high wavelength reproducibility and robustness, which become these filters suitable for industrial applications [31], [32].

The interferometers are the systems of wavelength selection of the Fourier Transformed NIR spectrometers (FT-NIR). The FT technology provides the possibility of describing any periodical signal in the time domain as a sum of sine or cosine signals with variable amplitudes and frequencies. Therefore, the sinusoidal electromagnetic waves of the NIR spectrum can be expressed using the FT. The interferometers generate a periodic signal at a scale of frequency under the audio-frequency (3.5 - 12 kHz), which can be easily related to the electromagnetic wave by means of the computational calculation of the FT algorithm, which generates the spectra in the frequency domain. The interferometer most often employed is the Michelson interferometer, which involves a beam splitter –commonly made of ZnSe or quartz- and two mirrors, one fixed and another moving. The two mirrors are initially equidistant to the beam splitter, then, each one produces a beam with roughly half of the intensity of the radiation emitted by the radiation source. These two beams will recombine at the beam splitter position after been reflected by the mirrors by means of a constructive interference, due to both will be in phase [24]. When the moving mirror is shifted toward (-) or away (+) from this position – this shift is called retard, δ - the polychromatic radiation undergoes destructive or partially destructive interference as functions of their wavelengths and the related distance. Therefore, the detector will register a signal with lower intensity. If the moving mirror is shifted by ± δ, regarding the equidistant point, at a constant velocity, this retard will produce a periodic signal modulated by the wavelengths of the beam as function of time, which is known as interferogram. For recovering the intensities associated to each frequency (or wavelength), the FT is applied to such resulting interferogram.

This process raises three relevant specific strengths of the interferometers:

- The Fellgett gain, which is associated to the intrisic gain in the signal to noise ratio due to the same total measurement time of the FT instrument, which will be sampling the same wavelenght *n* time more than an instrument based on a dispersive design. Because of the Fellgett gain, each measurement of the interferogram is in fact a simultaneous measurement of all wavelengths in the spectral region, which represents a gain regarding a dispersive system of wavelength selection.
- The Jacquinot gain, which is due to the absence of slits in the interferometer, and results in a high power reaching the detector.
- The Connes advantage, which is based on the use of a laser source with highly accurate and precise wavelength, for ensuring the periodic sampling of the interferometer. The Connes advantages is produced because of the monochromatic radiation of the laser (usually a HeNe laser) that allows the generation of a sinusoidal signal with a fixed frequency, at which the data acquisition of the main interferogram, coming from the NIR radiation, been retarded for changing and controlling the rate of the moving mirror [24].

NIR spectrophotometers based on interferometers and FT technology combine most of the best characteristics in terms of wavelength precision and accuracy, high signal-to-noise ratio and scan speed. Their capability of recording intensities of individual wavelengths in the NIR region enhanced their performance even regarding instruments with AOTF [33].

Finally, it is worth mentioning the Hadamard masks, which are spatial modulators employed for encoding the dispersed radiation and that can be used to improve the signal to noise ratio of dispersive instruments, taking advantage of the Fellgett gain. These spatial modulators can be of diverse forms. The most commonly employed are the linear encoders, which consist of a sheet of metal containing the slit patter according with a Hadamard matrix, driven by a step motor that empowers the encoding of the radiation. The encoded radiation can be later de-dispersed and delivered to a single detector by means of a Hadamard transform. The Hadamard transform is an algorithm mathematically simpler and computationally less demanding than the Fourier algorithm [34].

1.4.3 Sample chamber

Because of the wide variety of states of the matter and presentations in which the samples can be analysed using the NIRS, including liquids of diverse viscosities, solids of different sizes and shapes, there are diverse accessories that can be used as sample chambers. These can be fixed into the spectrometers or coupled to them. The main goal of all of these devices is always the proper spectrum acquisition with the minimum handling of the sample.

Some relevant examples of these parts of the NIR spectrometers, that are employed for the analysis of solid samples are the cubets, reflection probes for in-line configurations, custom-built tablet holders for the direct analysis of pharmaceutical tablets, among others. For in-line configurations, the sample chamber can be suppressed by the disposition of the other parts of the instrument (radiation source and detectors). Figure 8, illustrates some examples of devices developed for off-line (6A) and in-line (6B) analysis of solid samples.

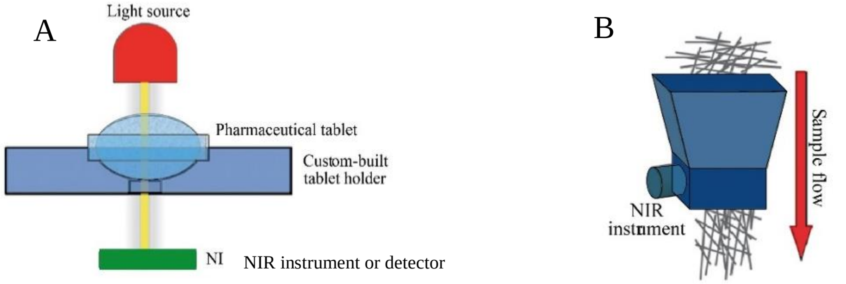


Figure 8. Illustration of devices for the presentation of solid samples to the NIR instrument: ***A*** custom-built tablet holder for pharmaceutical tablets analysed off-line; ***B*** reflection probe for in-line monitoring of solid products. Adapted from reference [15]

The cubets are generally made of quartz, material that is transparent to the IR radiation. Cubets are useful for analysing both solid and liquid samples. Additionally, for liquid samples, diverse accesories have been designed for transflectance measurements. These accessories can locate the sample in quarz cubets with a reflectance surface in one of the faces or can also be based on fiber probes as element of transmission of the incident radiation and reflected radiation. The reflection is produced in these cases by a reflectant surface located at known pathlength of the radiation out. Fiber probes properly set and fixed, allow fast, direct and reproducible NIR acquisitions of liquid and semiliquid samples.

1.4.4 Detector

The detectors employed in NIRS are photoelectric. In these detectors, the incident photons affect directly to the electronic state of the photosensible material employed for their construction, generating an electrical signal, which is the detector answer. The more frequently used is the detector of Lead sulfide (PbS), material that is a semiconductor with a proper sensibility in the region 1100 – 2500 nm at environmental temperature. For measurements under 1100 nm, detectors made of silicon provide a better performance.

Another kind of detectors are the Focal Point Array (FPA), which are equivalent in the NIR region of the Charged-Coupled Devices (CCD), used in the UV spectral region. These multichanel detectors allow faster recordings and enhanced signal to noise ratio regarding monocanal detectors. However, the use of these kind of detectors is restricted by the fact that their cost is so high as the cost of the spectrometer itself [31], [35].

A relevant issue regarding the detectors is the disposition in which they are used. For transmitance measurements, it is enough locating the detector aligned with the sample and the radiation source. On the other hand, for reflectance measurements, particularly of solid samples, the disposition of the detector can be modified as much as necessary for optimizing the caption of most of the radiation reflected by the sample. Figure 9, illustrates an example of the disposition of detectors for diffuse reflection measurements.

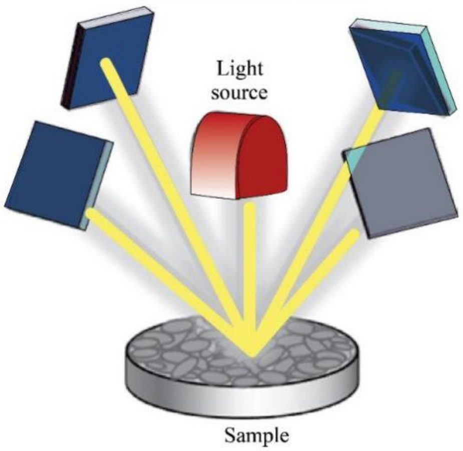


Figure 9. Disposition of detectors for Diffuse reflectance measurements. Adapted from reference [15].

For hand held instruments diverse spectral ranges become reachable, according to the kind of detector employed. Single point, arrays detectors of one and two dimensions can provide NIR spectra based on diverse vibrations depending on the material that constitute them. The InGaAs detectors – significantly more expensive than silicon-based detectors-, employed for Visible or Short wave NIR (400 – 1050 nm), provide only vibrational overtones (1000 – 1700 nm). Detectors made of extended InGaAs detectors, provide vibrational overtones and combinations bands (1200-2500 nm), however require the adaptation of a cooling system [14].

1.5 Data analysis

The interpretation of raw NIRS data is not straightforward from its visual inspection. This is the reason why using NIRS data implies the use of Chemometrics. Chemometrics can be understood as a discipline based on mathematic and statistic tools for collecting and interpreting information from chemical systems. Research in Chemometrics denotes diverse methods applicable to chemistry. There are tools for the design of experiments, for obtaining useful data from complex systems, optimizing experimental parameters, calibrating, signal processing, modelling and predicting structure-property relationships, for pattern recognition, among others [10].

The concept of Chemometrics was born from applications of mathematical statistics to problems of diverse scientific fields, as well as other areas as manufacturing and politics. The Chemometrics was prompted by advances in commercial processors in 1970, and it has positioned in the last years as a tool each time more important in Chemistry. Analytical Chemistry is from the very beginning the branch where its impact has been more evident.

According to Professor Luc Massart (Universiteit Brussel, Belgium), the starting date of Chemometrics depends on how it is defined. From his point of view, the article published by Bruce Kowalski in 1972, using Principal Component Analysis (PCA) for handling archeological data is the oldest one related to the modern chemometrical movement, even when it is possible to find previous findings that could be related to the area. For example, articles in univariate regression, confidence limits and all the types of regression methods.

According to Professor Svante Wold, (Umeå University, Sweden), with the "Student t-test" published by Gosset en 1908, took place the first article of Chemometrics. However, as it is known today, with the capability of handling huge amounts of data from many diverse chemical problems, started at the end of 1960, with publications of Malinowski, Kowalski, Eisenhour and Jurs, all of them in Analytical Chemistry.

For Professor Bruce Kowalski (University of Washington, USA), it is hard to define an event in a particular time for remarking such starting point. Nevertheless, he considers the creation of the Chemometrics Society as a major issue in this sense. Based on that, he opines that Chemometrics was constituted actually as a field of study when the Chemometrics Society was formed, on June 10th, 1974 [36].

Professors Massart, Wold and Kowalski were fundamental actors of the starting of Chemometrics [37]–[44]. This is the reason why their views have been considered relevant insights for the topic.

Wold, who proposed the term Chemometrics in 1972, suggests that from a philosophical point of view it is always important to emphasize that a huge part of chemometrical methodology is based on a deductive vision of science, in the "indirect method of research" [44]. It is like this because statistics, as all the formal sciences, has as its own method the deductive one, which is evidenced not only in this application to Chemometrics, but in other areas of chemoinformatics as well [40]. Professor Wold, with formation in Statistics, was a frequent reader of Biometrika and the Journal of Biometrics, long time before proposing the name of Chemometrics. Because of that, it is easy to understand that the name comes from biometrics or psychometrics. In this sense, it is useful to highlight that Partial Least Squares regression (PLS), a tool often used nowadays in Chemometrics, was described by P. Horst in 1961 in Psychometrica, where the same author indicated that this kind of problems and solutions had been already presented by Hotelling in previous reports [36].

Conceptually, biology has a bigger amount of data than chemistry, but smaller than psychology. Therefore, psychology had probably the hardest work at the very beggining and that is why Biometrics came later. However, why Chemometrics did not come after Psychometrics instead of Biometrics still looks as an interesting question.

Nevertheless, what is clear is that Psychometrics and Biometrics are more qualitative theories than Chemometrics. Additionally, it is possible to find a relation with the need of quantitative results in the industry, which –along with academy- has provide valuable contributions to the development of Chemometrics. Indeed, the industry was the natural place for the application of Chemometrics and is still one of its more important drivers [38].

One evidence of the industrial support in the development of Chemometrics is that the more important windows for sharing knowledge in this field –besides the Journal of Chemometrics and Chemometrics and Intelligent Laboratories Systems, bulletins founded and sponsored by the Chemometrics Society- are journals focused on industrial applications of Analytical Chemistry. Some examples are: Analytica Chimica Acta, Analytical Chemistry, Applied Spectroscopy, Technometrics, Journal of Chromatography and the Journal of Chromatography. Recently, also specialized journals of pharmaceutics, food, petrochemical, environmental, forensic, among others areas are important communication means in this field.

Currently, the International Chemometrics Society is constituted by representatives from Chemometrics Societies from Sweden, Spain, Russia, Norway, Italy, Germany, France, Findland, The Netherlands, Denmark, Czech Republic, United Kingdom, Belgium, South Africa, North America and Australia. This organization has provided the following definition:

> *“Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.”*

Figure 10, shows the relevant events regarding the historical development of Chemometrics.

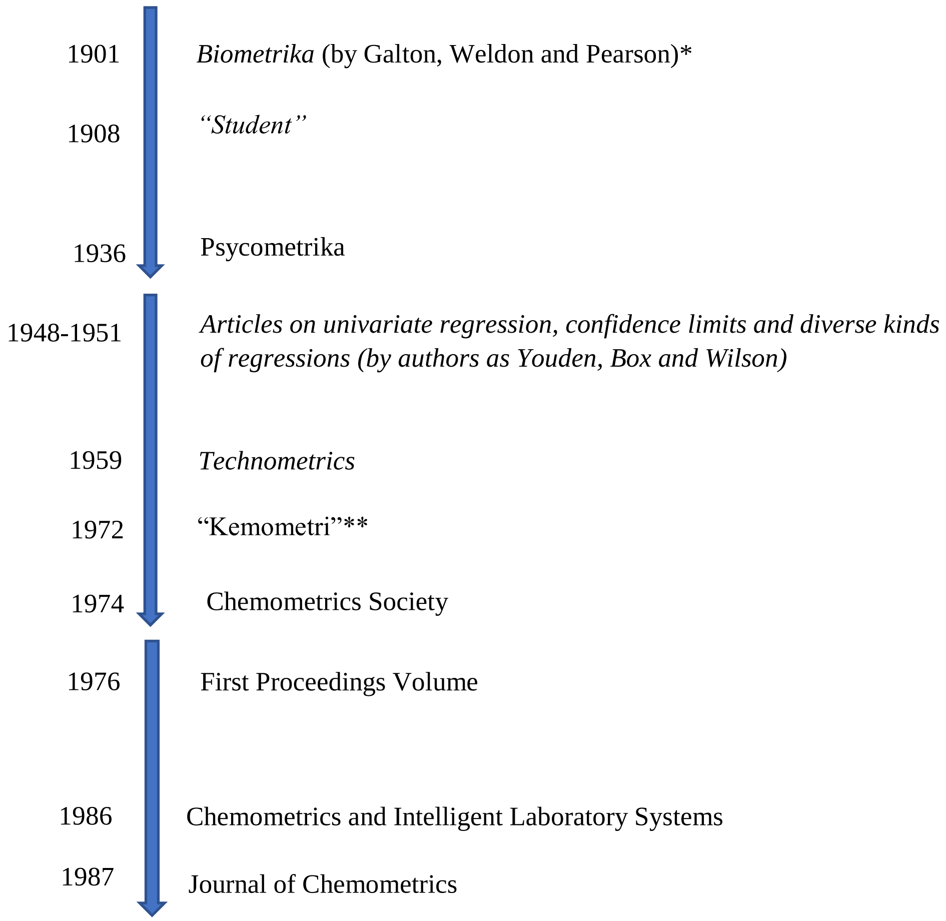

1901
Biometrika (by Galton, Weldon and Pearson)*
1908
"Student"
1936
Psycometrika
1948-1951
Articles on univariate regression, confidence limits and diverse kinds of regressions (by authors as Youden, Box and Wilson)
1959
Technometrics
1972
"Kemometri"**
1974
Chemometrics Society
1976
First Proceedings Volume
1986
Chemometrics and Intelligent Laboratory Systems
1987
Journal of Chemometrics

Figure 10. Summary of relevant events in the gestation of Chemometrics in a non-lineal scale. Adapted from scheme published by Geladi y Esbensen [37].

**Francis Galton, Walter Weldon and Karl Pearson were Britain Scientifics that found Biometrics as a common interest. All they three accomplished relevant contributions to scientific knowledge and institutions of the XX century. One of them was the first Department of Statistics of the Word, founded by Pearson at the University College of London in 1911 [11].*

***Due to native language of Professor Wold was Swedish, the original expression proposed by him was Kemometri, which was later translated to English as Chemometrics.*

Multivariate data analysis is the core of Chemometrics, even when the discipline covers a wider range of tools. Most of the initial papers in the literature in the 1970s used simple multivariate methods (based on univariate methods) to explore complex chemical data sets. Traditionally, one or two variables (e.g. the intensity at characteristic wavelengths) were used to characterize a sample. However, this procedure had the drawback of requiring the selection of discreet variables, which often results in the loss of information and the use of information mixed up in the signal.

The increment of the number of variables that spectroscopic techniques, such as NIRS, provide, lead to difficulties when using univariate methods. This is the reason why the multivariate data analysis constitutes an indispensable tool for the NIR spectra. The following sections describe the general concepts currently accepted for the particular tools of multivariate analysis employed for the development of the research presented in this thesis.

1.5.1 Stages in the development of multivariate analytical methods based on NIRS data

In general, both qualitative and quantitative analytical methods based on NIRS data require the establishment of models for generating answers and results. The development of such models involves several stages that must be completed before applying the models to data of new and unknown samples. These stages are:

a) Selection of the calibration set

It is required the availability of samples representatives of all the sources of variability that will be entailed in the process or system to be studied. Sources of variability are all those parameters that can change their magnitudes during the process, generating diverse values or spectral characteristics over the range considered. The sources of variability have to be evaluated in each particular case, due to they can be very diverse, depending on each analytical case.

b) Data acquisition

The proper sampling conditions, presentation of the samples to the spectrometer, acquisition mode and instrument configuration need to be evaluated according to each particular case. Chemical and physical characteristics of the analyte, the matrix and the intended purpose of the method must be stated at this point. Once these settings are established, the data recording can be completed.

c) Visualization of the data

Before the application of any mathematical or statistical tool, it is fundamental to ensure the quality of the NIRS data to be employed. Visualization with and without data pre-treatments is useful to identify erroneous recordings, NIR spectra of samples non representative of the range of variability to be studied, among other incidences that must be eliminated from the set that is intended to be used during the calibration of the model. These kind of spectra, which are no representative of the samples to be studied, are named outliers. Ensuring clear and reliable data can save valuable time and efforts during modelling. For very complex data sets, this step can also require the use of exploratory data analysis tools.

d) Establishment of reference parameters

Multivariate analytical methods based on NIRS data are always founded on reference data. This is the reason why NIRS methods are secondary methods from the analytical point of view, due to they require data from concentration methods –also known as primary methods- to calculate their results. Therefore, the accuracy and precision of reference methods is quite important for the development of new methods based on NIRS. The reference parameters are the starting point of the analytical performance of the new method, even when it can be awared that precision can be improved using NIRS but accuracy not.

e) Data pre-treatment

The use of spectral pre-treatments is most of the times a necessary stage during the development of new methods. The objectives of the data pre-treatment are increasing the signal to noise ratio and removing information not relevant for modelling and prediction with the NIRS data. The pre-treatments are mathematical transformations that allow taking advantage as much as possible of chemical and physical information of interest from the studied system and eliminating systematic errors of the measurements, such as nonlinear instrument responses, shift problems, scattering effects, and interfering chemical and physical variations [45]. Each mathematical pre-treatment provides a particular outcome on the NIRS data, and their purposes and strategies are diverse. Furthermore, the order in which they are applied have an effect on the final result [46]. Even when there are general descriptions of the intended purpose of each pre-treatment, the selection of the proper pre-treatment(s) and the appropriate combination of them is basically an empiric process that needs to be accomplished considering each particular NIRS data set.

The application of mathematical pre-treatments, as well as all the chemometric tools, is enabled by the description of NIRS data in matrices. In these matrices the samples are defined in the rows and the wavelengths of the spectra in the columns. The pre-treatments used in this thesis can be classified in terms of the sense in which they can be calculated, as:

Pre-treatments calculated along the rows of samples:

<u>Standard Normal Variate (SNV):</u> Useful for correcting the scattering effect due to the differences of particle size of solid samples [47]. This pre-treatment is based on the correction of the spectrum regarding its standard deviation, principle that it has in common with the Normalization and the Multivariate Scattering Correction (MSC). The general equation for the SNV pre-treatment can be described as:

$$x_{corrected} = \frac{x_{original} - a_0}{a_1} \qquad \textit{Equation 17}$$

where $x_{original}$ corresponds to each individual absorbance value of the spectrum. For SNV, a_0 is the average value of the spectrum to be corrected, however for Normalization this term is set equal to zero. For SNV, a_1 is the standard deviation of the spectrum of each sample [48]. Calculations of MSC involves additionally the average standard deviation and the grand mean of the spectrum, which generates results generally similar to SNV, however MSC results can variate depending on the characteristics of the data set [49].

<u>Derivatives:</u> Due to the derivative of a constant number is equal to zero, the use of derivative on NIRS data allow to emphasize the bands where the main differences between the spectra can be found. Derivatives remove both additive and multiplicative effects: the first derivative eliminates the baseline shifts and the second one eliminates the terms that variate lineally with the wavelength. In the most basic method for derivation, the first derivative is calculated as the slope between two subsequent point of the spectrum. At that point, the second derivative is estimated by the slope between two successive points of the first-order derivative spectrum. However, when this method is directly applied to NIRS data, it produces noise inflation. This is the reason why different approaches have been developed to avoid the noise inflation in finite differences [48]. The approaches most commonly employed were developed by Norris-Williams [50] and by Savitzky-Golay (more complex algorithm) [51]. Both procedures use smoothing to ensure not reducing too much the signal-to-noise ratio in the corrected spectra.

<u>Ortogonal Signal Correction (OSC):</u> The pre-treatments previously described may remove information from the NIRS data that could be correlated to the property to be determined. The orthogonal signal correction is a variant of a partial least squares regression that can be used to transform a data set as close to orthogonal as possible to a given the expected response. This

procedure ensures that the signal correction removes as little information as possible regarding the reference values [52]. However, it has to be applied with precaution, because it can generate overoptimistic results.

Pre-treatments calculated along the columns of wavelengths:

Averaging: Based on the calculation of the average value of the intensities of diverse spectra, with the objective of reducing the random noise and/or resolution.

Mean centring: This is a procedure completed for centring the data regarding each particular set. This pre-treatment comprises the calculation of the mean value at each wavelength of the calibration set, and the further subtraction of such mean from each point of the corresponding column [49]. After this pre-treatment, the mean value is considered the centre of the model and all the variables are referred to this centre, keeping the original units.

Scaling: After the mean centring, the resulting values per column can be divided by the standard deviation of each wavelength. This calculation allows that the variance of each variable has the value of the unit.

f) Modelling (construction of the calibration model)

This stage involves the selection of the parameters and acceptance criteria that describes the relation between the analytical signal and the reference value expected. The selected features state the called calibration model, which can be empiric or can also be clearly explained for a theoretical fundamental. The optimization of the model is achieved by testing diverse chemometrical algorithms, data pre-treatments, wavelength ranges, among other variables.

g) Validation of the model

After selecting the proper parameters and acceptance criteria, the calibration model must be challenged by new samples, with known reference values, but not included in the calibration set of data. The prediction of the properties using these new samples are useful to determine the analytical features of the NIRS method. In general, the validation of qualitative models is described in terms of selectivity (specificity) and robustness. For quantitative models, the validation contemplates the linearity, the accuracy (at target level and over the range studied), the repeatability, intermediate precision and the robustness.

The predictive capability of quantitative models is often evaluated using the Root Mean Square Error of Prediction (RMSEP), which evaluates the response of the model based on NIRS data regarding the reference method. It can be calculated by:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i^{NIRS} - Y_i^{Ref})^2}{n}} \qquad \textit{Equation 18}$$

where *n* is the total number of samples, and Y^{NIRS} and Y^{Ref} are the magnitudes of the properties predicted by the model and the reference method respectively.

1.5.2 Qualitative analysis

The qualitative analysis of NIRS data is based on the comparison of new spectra to spectra previously established as reference data. The chemometrical methods employed for such comparison are named Pattern Recognition Methods (PRM). The PRM are based on mathematic calculations of correlations or distances. Generally, the PRM can be classified as no-supervised and supervised. In this thesis, the no supervised method used has been the PCA. The supervised methods employed have been the correlation, the Euclidean and Mahalanobis distances and the discriminant analysis.

a) Non-supervised methods

The exploratory data analysis (EDA) of NIRS data, as well as diverse qualitative and quantitative chemometrical methods, take advantage of the Principal Components Analysis (PCA). The PCA is probably the most widely known multivariate chemometric technique, because of its usefulness for reducing the number of variable of complex data sets and visualizing in a space of 2 or 3 dimensions, the similarities and differences between unknown samples, and visualization of loadings.

The PCA relates a data matrix to a number of factors, which are calculated taking into account the variance of the data set. The interpretation of results of PCA for classification purposes can be done by means of the representation of the scores of the samples of a principal component (PC) versus the scores of another PC. This graphical representation is known as the scores plot. If there is a relationship between the samples, in the scores plot groups of points that can be correlated with one or more characteristics of the samples will be displayed [53]. Each PC is described not only in terms of the scores of the samples, but also involves information in the loadings, which are related to the NIRS spectra. The loadings are useful to understand the ranges of the spectra that have influence on the studied data set.

Before applying a PCA to NIRS spectra, it is required to complete a mean centring pre-treatment of the data, in order to expose systematic variation in variables with small impact in the structure of the data and to retain those variables which are more relevant. This transformation also makes the distribution of each variable more normal [54]. Afterwards, the samples undergo a mathematical transformation of the original data matrix, $\boldsymbol{X}$, based on the relation:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{t}} + \boldsymbol{E} \qquad \textit{Equation 19}$$

where $\boldsymbol{T}$ represents the scores and have the same number of rows of the original data matrix (the total number of samples). $\boldsymbol{P}^{t}$ are the loadings, and have the same number of wavelengths of the original data matrix, and $\boldsymbol{E}$ is the error or noise involved in the matrix.

b) Supervised methods

These methods require the initial definition of classes or set of data to which new samples should belong, i.e. regarding to which new data have to be compared. For the use of these kinds of methods it is necessary the availability of a data set representative of the class that is required to identify. This data is used as reference for the comparison of new samples. Some supervised methods are:

Correlation: The identification is completed by the calculation of the correlation coefficient between the spectrum of the new sample and the average spectrum of the class defined during modelling. It is required to establish a correlation coefficient threshold in order to evaluate the identification results after the calibration of the qualitative model [22].

Distances: There are other kind of comparisons that are based on the calculation of the distance that represents how different is a sample regarding to another or regarding the point of the space that represents a particular class in the model. The Euclidean distance is calculated on the space of the wavelengths. In this method, each class is described as a hypersphere, whit a fixed radio. A new sample will be identified as belonging to a particular class if it is inside of such radio.

The Mahalanobis distance is calculated using the covariance matrix. Similarly, the class is defined during calibration, but by means of a as an ellipse described from a PCA. The identification of new samples is evaluated with respect to the distance to the centre of the ellipse. The main difference between the Euclidean and the Mahalanobis distances is that the Mahalanobis distance includes the covariance term in its calculation [55].

The most relevant application of supervised methods to NIRS data in this thesis has been related to the construction of Spectral libraries. Spectral libraries based on NIRS data are considered supervised PRM useful to create classes from set of data characterized by reference techniques. The identification of new samples is achieved by the comparison of their spectra with those of all the classes that constitute the library. Each library can involve diverse qualitative objectives across the whole set of classes, which discrimination can require diverse strategies. The selection of such strategies is based on the particular characteristics of each data set.

Discriminant analysis: These methods are based on discriminant functions that divide the space in characteristic regions for each class, creating boundaries between each one of them. The discriminant methods most commonly applied to NIRS data are the Lineal Discriminant Analysis (LDA) and the Quadratic Discriminant Analysis (QDA) [55].

1.5.3 Quantitative analysis

The quantitative analysis of NIRS data allows to relate the instrumental answer and the chemical or physical property to be determined. For this purpose, it is required a representative set of samples and reliable reference data. NIRS can provide simultaneously information from diverse analytes. This fact has driven the development of calibration methods able to relate multiple variables to the property of interest. These are known as the multivariate calibration methods.

These methods can be classified in terms of their characteristics in lineal and non-lineal methods. In the non-lineal methods, the Artificial Neural Networks (ANN) and the non-lineal Partial Least Squares (non-lineal PLS) can be found. In the lineal methods, there are tools for completing the calculations using the original variables, as the Multivariate Lineal Regression (MLR) and tools for calculations based on the reduction of the variability of the data sets. In this last group, the Principal Components Regression (PCR) and the Partial Least Squares (PLS) regression can be termed.

In this thesis, the multivariate calibration method employed for quantitative purposes was the PLS regression. This strategy was introduced by Wold in 1975, and it has been labelled as “soft modelling” due to it does not made a priori assumptions about the model structure. The PLS approach is useful for calibrating complex spectral NIR data with interference effects from other factors than those related to the analyte(s) [56].

The PLS regression makes use of the reduction of variability considering both the NIRS data set, ***X*,** and the known values of the property studied (obtained from the reference method), ***Y***. The objective of this procedure is including most of the relevant information for the prediction of properties of new samples in the first components. Because of that, the PLS regression uses both the NIRS data and the reference data for the description of new variables, named latent variables, factors or components. The first step is also mean centring the data. Then, each of the matrices is described in terms of A<K, where K is the number of original variables of the matrix **X**. This makes possible to calculate simultaneously components by:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} \qquad \textit{Equation 20}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\boldsymbol{T}} + \boldsymbol{F} \qquad \textit{Equation 21}$$

where the ***T*** is the scores matrix, ***P*** the loadings, and ***E*** the residuals matrices for the NIRS data; and the ***U*** is the scores matrix, ***Q*** the loadings, and ***F*** the residuals matrices for the concentrations or properties obtained considering the reference method. Therefore, the loadings in the PLS does no express only the maximum variability of the samples in the NIR spectra, because they are corrected to obtain the maximum predictive capability for the matrix ***Y***. In cases of calculation of only one concentration or property of the matrix ***Y***, the algorithm is indicated as PLS1, which can be considered as a simplification of the general algorithm, known as PLS2 [56].

In the PLS regressions of NIRS data presented in this thesis, the condition of homocedasticity has been assumed. The assumption of homocedasticity implies the consideration of non-changes in the variance occur across all values of the independent variables (wavelengths). This assumption is common to linear regression models calculated using NIRS data. The heterocedasticity is the violation of homocedasticity and its presence in data employed for regression models impacts the minimization of the residuals, which is the main goal of a PLS regression, the minimum residuals as possible. Some strategies can be employed for dealing with the heterocedasticity of data, such as weighted least squares regressions. However, it is important to take into account that these calculations also require additional assumptions [57].

1.6 General advantages and disadvantages of the technique

At this point, it is worth remaking the general advantages and challenges of the NIRS technique that can be defined from the states described in this chapter.

As **advantages** can be outlined the following:

- Is a non-destructive and non-invasive technique
- Sample preparation is too simple and sometimes even no required. Minimum sample handing both for liquid and solid samples, which allow a huge number of analysis of samples for Quality Control purposes.
- Each analysis represents a very low cost, due to the absence of reactives, solvents among other materials required for the application of most of the concentration techniques. This fact increases the analytical capability of the laboratory.
- The technique enables the analysis of diverse analytes with the same spectral acquisition, which allows the automatization of several activities.
- Simultaneous evaluation of physical and chemical parameters of the samples.
- Since the detection system is simple and have no moving parts, is a technique particularly useful for process control procedures.
- In most of the fields of application of the technique, the accuracy of the technique is comparable to the reference techniques, and generally the precision is even better due to the no sample preparation requirement.

Even though, as every analytical technique, the NIRS also faces **challenges**:

- The complexity of the NIR signal requires the use of chemometric techniqes for modelling the data before identifying and/or quantifying new unknown samples.
- The calibration process can be challenging, since it is necessary the availability of samples for increasing the concentration range as well as other physical and chemical sources of variability.
- It is not possible to analyse new unknown samples with sources of variability different to those considered during the calibration.
- The NIRS provide low sensibility, particularly in diffuse reflectance measurements. This fact, in general, constrains the analysis of low concentration analytes.

- Since the effect of small optical and electronic differences among similar instruments, the transference of calibrations between them is not straightforward.

1.7 Applications of NIRS

Based on the advantages and despite the drawbacks previously described, nowadays the NIRS is considered a mature technique, broadly applied in the most diverse fields [58]. The first and more traditionally encompassed field of application is the food and agro products. Some of the applications in agriculture are related -but not restricted- to the classification and quantification of different analytes in coffee, wine, fruits, milk, meat, honey, cotton, natural products, cheese, olive oil, potatoes, fish, among other applications regarding transgenic cultivations, food safety and contaminants [58]. In the fuel industry there also diverse applications particularly interesting from the refining perspective, such as predictions of chemical and physical properties of crude oil, gasoline, diesel, naphtha, among applications to biofuels [59].

In the production of polymers diverse methods for the synthesis monitoring, among other general applications can be found [60]. The study of soils has been also benefited by the use of the NIRS, which has been employed for estimating diverse characteristics, such as contaminants and organic matter content [61]. Applications of NIRS have been also developed in the field of forestry [62], wood [63] and paper [64], environmental analysis, including even the wildlife and biodiversity research, area in which the NIRS is employed to embrace the biodiversity by means of a broad range of ecological and evolutionary analysis [65].

Among other fields where NIRS has been able to provide useful answers and solutions, it is possible to find the three areas that has been included in the scope of this thesis: control of New Psychoactive Substances (NPS), biotechnology and pharmaceutics. Antecedents of the technique related to the fields of NPS control and biotechnology are included in the introduction of chapters 3 and 4, respectively.

On the other hand, a description of the applications of the technique to the field of pharmaceutics has been included as part of the present introductory chapter. It is important to highlight that the following section, 1.6.1, (pages 50 to 65) has been prepared as a contribution to the book ***Introduction to Process Analytics for Pharmaceuticals***, belonging to the collection ***Advances in***

Pharmaceutical Technology, ISBN-10: 1119433029. This book is Edited by J. Rantanen et al, and it will be published by Wiley-Blackwell in 2019.

1.7.1 NIRS applied to the Pharmaceutical industry

Near Infrared Spectroscopy (NIRS) is an analytical technique based on the interaction between electromagnetic radiations in the wavelength range 780-2500 nm and matter. The NIR spectrum can be considered as a "*fingerprint*" of each chemical compound or mixture of them that is analysed, which contains absorption bands that are the result of overtones and combinations of the fundamental vibrations observed in the Mid-Infrared (MIR) region (wavelength range 2500-6000 nm). Depending on the location of the detectors regarding the sample, NIR spectra can be acquired in reflectance, transmittance and transflectance modes. In general, reflectance detection mode is used for solid samples and transmittance and transflectance modes for liquids and gases.

Additionally, NIRS of solids is sensible to the scattering effect caused by physical characteristics of the samples, such as particle size, compaction and polymorphism. This is the reason why it can provide not only chemical but also physical information from the same analysis. Therefore, even when NIR spectra show broad and overlapped bands –which makes necessary the use of Chemometrics for extracting the proper information from them- its simultaneous sensitivity to chemical and physical conditions becomes it into an effective process analytical tool for many industrial fields. The use of Chemometrics implies the application of statistical and mathematical methods to chemical data, with the aim of achieving the maximal collection and extraction of useful information from it [66], therefore, its application to NIRS experimental works profits from the information obtained from them. One of the most representative examples of the capability of the combination of NIRS and Chemometrics, is provided by the NIRS applications to the pharmaceutical industry.

In fact, NIRS has been successfully applied to all the stages of pharmaceutical manufacturing, from the very original raw materials to finished products, where the value of its advantageous features compared to other concentration techniques has been widely demonstrated [67], [68]. The most relevant of these advantages are its very short response time, non-sample preparation requirement and non-generation of residual solvents. Furthermore, the availability of fibre-optic probes of diverse sizes and characteristics, allows the acquisition of NIR spectra from process units

1.8 References

[1] Renneboog, R. M. J., and MSc, “International System Of Units.,” *Salem Press Encycl. Sci.*

[2] L. De Galan, “The four players in the analytical performance,” *J. Anal. At. Spectrom.*, vol. 27, no. 8, pp. 1173–1176, 2012.

[3] R. Belcher and author, “The Role of Analytical Chemistry in Academic and Industrial Chemistry [and Discussion],” *Philos. Trans. R. Soc. London. Ser. A, Math. Phys. Sci.*, vol. 305, no. 1491, 1982.

[4] E. F. J. Ring, “The discovery of infrared radiation in 1800,” *Imaging Sci. J.*, vol. 48, no. 01, pp. 1–8, 2000.

[5] W. Herschel, “Experiments on the Refrangibility of the invisible Rays of the Sun,” *Philos. Trans. R. Soc. London*, vol. 90, no. 3, pp. 284–292, 1800.

[6] Sheposh and Richard, “Electromagnetic spectrum.,” *Salem Press Encyclopedia of Science*. 2017.

[7] Y. M. Rabkin, “Technoological Innovation In Science: The Adoption of Infrared Spectroscopy by Chemists,” *Isis*, vol. 78, no. 1, pp. 31–54, 1987.

[8] F. Rose, “Quantitative analysis, with respect to the component structural groups, of the infrared (1 to 2 M) molal absorptive indices of 55 hydrocarbons,” *J. Res. Nat. Bur. Stand.*, vol. 20, no. 2, p. 129, 1938.

[9] H. Hall, B. E. Simpson, and H. A. Mottola, “Design and performance of a ‘titration head’ and special cell holder for the cary 14 spectrophotometer,” *Anal. Biochem.*, vol. 45, no. 2, pp. 453–461, Feb. 1972.

[10] G. L. Bosco, “James L. Waters Symposium 2009 on near-infrared spectroscopy,” *TrAC - Trends Anal. Chem.*, vol. 29, no. 3, pp. 197–208, 2010.

[11] J. M. Conway, K. H. Norris, and C. E. Bodwell, “A new approach for the estimation of body composition: Infrared interactance,” *Am. J. Clin. Nutr.*, vol. 40, no. 6, pp. 1123–1130, 1984.

[12] J. B. Callis, D. L. Illman, and B. R. Kowalski, “Process analytical chemistry,” *Anal. Chem.*, vol. 59, no. 9, pp. 624–637, 1987.

[13] J. Workman, B. Lavine, R. Chrisman, and M. Koch, “Process Analytical Chemistry,” *Anal. Chem.*, vol. 83, no. 12, pp. 4557–4578, Jun. 2011.

[14] R. Crocombe, “Portable Spectroscopy,” *Appl. Spectrosc.*, vol. 72, no. 12, pp. 1701–1751, 2018.

[15] M. Kutz, *Handbook of Measurement in Science and Engineering*, vol. 3. Chichester, West Sussex: Wiley, 2016.

[16] B. G. Osborne and T. Fearn, *Near infrared spectroscopy in food analysis*. Longman Scientific & Technical, 1986.

[17] H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise, Eds., *Near-infrared spectroscopy*, First edit. Weinheim: Wiley-VCH, 2005.

[18] Y. Ozaki, M. Ishigaki, Y. Futami, and C. W. Huck, “Introduction of Quantum Chemical Calculation

for near Infrared Spectroscopy," *NIR news*, vol. 27, no. 7, pp. 8–11, Oct. 2016.

[19] M. A. Czarnecki, Y. Morisawa, Y. Futami, and Y. Ozaki, "Advances in Molecular Structure and Interaction Studies Using Near-Infrared Spectroscopy," *Chem. Rev.*, vol. 115, no. 18, pp. 9707–9744, 2015.

[20] R. C. Denney, *A dictionary of spectroscopy*, Second. Minnesota: Wiley, 1982.

[21] D. A. Skoog, F. J. Holler, and S. R. Crouch, *Principles of instrumental analysis*, Seventh ed. Boston: Cengage Learning, 2018.

[22] D. A. Burns and E. W. Ciurczak, *Handbook of Near-Infrared Analysis*, Third edit. New York: CRC Press, 2008.

[23] B. D. Guenther, *Modern optics*, Second edi. Oxford: OUP Oxford, 2015.

[24] M. Alcalà *et al.*, "Near-infrared Spectroscopy in Laboratory and Process Analysis," *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, pp. 1–39, 2012.

[25] A. S. Bonanno and P. R. Griffiths, "Discrimination of Organic Solvents Using an Infrared-Emitting Diode-Based Analyzer. Part I: Feasibility," *Appl. Spectrosc.*, vol. 49, no. 11, pp. 1590–1597, Nov. 1995.

[26] A. S. Bonanno and P. R. Griffiths, "Discrimination of Organic Solvents Using an Infrared-Emitting Diode-Based Analyzer. Part II: Practical Demonstration," *Appl. Spectrosc.*, vol. 49, no. 11, pp. 1598–1607, Nov. 1995.

[27] S. Prucnal, K. Gao, W. Anwand, M. Helm, W. Skorupa, and S. Zhou, "Temperature stable 13 µm emission from GaAs," *Opt. Express*, vol. 20, no. 23, p. 26075, 2013.

[28] H. H. (Hobart H. Willard and A. Rojas Hernández, *Métodos instrumentales de análisis*. México : Grupo Editorial Iberoamérica, 1991.

[29] H. A. (Hugh A. Macleod, *Thin-film optical filters*. .

[30] W. A. Atia, D. C. Flanders, P. Kotidis, and M. E. Kuznetsov, "MEMS Fabry Perot filter for integrated spectroscopy system," 28-Apr-2006.

[31] P. J. Treado, I. W. Levin, and E. N. Lewis, "Near-Infrared Acousto-Optic Filtered Spectroscopic Microscopy: A Solid-State Approach to Chemical Imaging," *Appl. Spectrosc.*, vol. 46, no. 4, pp. 553–559, Apr. 1992.

[32] C. D. Tran, "Acousto-Optic Devices," *Anal. Chem.*, vol. 64, no. 20, pp. 971A-981A, Oct. 1992.

[33] C. Pasquini, "Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications," *J. Braz. Chem. Soc.*, vol. 14, no. 2, pp. 198–219, 2003.

[34] M. Alcalà *et al.*, "Near-Infrared Spectroscopy in Laboratory and Process Analysis," in *Encyclopedia of Analytical Chemistry*, Chichester, UK: John Wiley & Sons, Ltd, 2012.

[35] Q. S. Hanley, C. W. Earle, F. M. Pennebaker, S. P. Madden, and M. B. Denton, "Peer Reviewed: Charge-Transfer Devices in Analytical Instrumentation," *Anal. Chem.*, vol. 68, no. 21, pp. 661A-667A, Nov. 1996.

[36] P. Geladi and K. Esbensen, “The start and early history of chemometrics: Selected interviews. Part 1,” *J. Chemom.*, vol. 4, no. 5, pp. 337–354, Sep. 1990.

[37] K. Esbensen and P. Geladi, “The start and early history of chemometrics: Selected interviews. Part 2,” *J. Chemom.*, vol. 4, no. 6, pp. 389–412, Nov. 1990.

[38] R. G. Brereton, “A short history of chemometrics: a personal view,” *J. Chemom.*, vol. 28, no. 10, pp. 749–760, Oct. 2014.

[39] J. A. Pérez-Bustamante, “A schematic overview of the historical evolution of Analytical Chemistry,” *Fresenius. J. Anal. Chem.*, vol. 357, no. 2, pp. 151–161, Jan. 1997.

[40] J. Gasteiger, “Chemoinformatics: a new field with a long tradition,” *Anal. Bioanal. Chem.*, vol. 384, no. 1, pp. 57–64, Jan. 2006.

[41] C. E. Miller, “Chemometrics in Process Analytical Technology (PAT),” in *Process Analytical Technology,* Chichester, UK: John Wiley & Sons, Ltd, 2010, pp. 353–438.

[42] G. Ramis Ramos and M. C. García Álvarez-Coque, *Quimiometría*. Madrid : Síntesis, 2001.

[43] M. Otto, *Chemometrics : statistics and computer application in analytical chemistry*. Wiley-VCH, 2017.

[44] S. Wold, “Chemometrics; what do we mean with it, and what do we want from it?,” *Chemom. Intell. Lab. Syst.*, vol. 30, no. 1, pp. 109–115, Nov. 1995.

[45] S. D. (Steven D. . Brown, L. A. Sarabia, and J. Trygg, *Comprehensive chemometrics : chemical and biochemical data analysis*. Elsevier, 2009.

[46] T. Fearn, “Are two pretreatments better than one?,” *NIR News*, vol. 14, no. 6. pp. 9–11, 2003.

[47] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, “Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra,” *Appl. Spectrosc.*, vol. 43, no. 5, pp. 772–777, Jul. 1989.

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

[49] D. Granato and G. Ares, *Mathematical and Statistical Methods in Food Science and Technology*, First. Chichester, United Kingdom: Willey Blackwell, 2014.

[50] N. K.H., “Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size.,” *Cereal Chem.*

[51] A. Savitzky and M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.,” *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[52] S. Wold, H. Antti, F. Lindgren, and J. Öhman, “Orthogonal signal correction of near-infrared spectra,” *Chemom. Intell. Lab. Syst.*, vol. 44, no. 1–2, pp. 175–185, Dec. 1998.

[53] R. G. Brereton, “Pattern recognition in chemometrics,” *Chemom. Intell. Lab. Syst.*, vol. 149, pp. 90–96, 2015.

[54] O. M. Khalheim, “Scaling of analytical data,” *Anal. Chim. Acta*, vol. 177, pp. 71–79, Jan. 1985.

[55] P. Gemperline, *Practical guide to chemometrics*. CRC/Taylor & Francis, 2006.

[56] R. Manne, “Analysis of two partial-least-squares algorithms for multivariate calibration,” *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 187–197, Aug. 1987.

[57] L. Cuadros-Rodriguez, A. González-Casado, A. M. García-Campaña, and J. L. Vílchez, “Ensuring both normality and homocedasticity of chromatographic data-ratios for internal-standard least-squares calibration,” *Chromatographia*, vol. 47, no. 9–10, pp. 550–556, May 1998.

[58] C. Pasquini, “Near infrared spectroscopy: A mature analytical technique with new perspectives – A review,” *Anal. Chim. Acta*, Apr. 2018.

[59] H. Chung, “Applications of Near-Infrared Spectroscopy in Refineries and Important Issues to Address,” *Appl. Spectrosc. Rev.*, vol. 42, no. 3, pp. 251–285, May 2007.

[60] M. Watari, “A Review of Online Real-Time Process Analyses of Melt-State Polymer Using the Near-Infrared Spectroscopy and Chemometrics,” *Appl. Spectrosc. Rev.*, vol. 49, no. 6, pp. 462–491, Aug. 2014.

[61] V. Genot, G. Colinet, L. Bock, D. Vanvyve, Y. Reusen, and P. Dardenne, “Near Infrared Reflectance Spectroscopy for Estimating Soil Characteristics Valuable in the Diagnosis of Soil Fertility,” *J. Near Infrared Spectrosc.*, vol. 19, no. 2, pp. 117–138, Apr. 2011.

[62] C.-L. So *et al.*, “Near Infrared Spectroscopy in the Forest Products Industry, Forest Products Journal,” *For. Prod. Journal, Vol. 54 No. 3. March 2004. p. 6-16*, 2004.

[63] S. Tsuchikawa and H. Kobori, “A review of recent application of near infrared spectroscopy to wood science and technology,” *J. Wood Sci.*, vol. 61, no. 3, pp. 213–220, 2015.

[64] T. Trung, G. Downes, R. Meder, and B. Allison, “Pulp mill and chemical recovery control with advanced analysers - from trees to final product,” *Appita*, vol. 68, no. 1, pp. 39–46, 2015.

[65] C. K. Vance, D. R. Tolleson, K. Kinoshita, J. Rodriguez, and W. J. Foley, “Near Infrared Spectroscopy in Wildlife and Biodiversity,” *J. Near Infrared Spectrosc.*, vol. 24, no. 1, pp. 1–25, Feb. 2016.

[66] P. K. Hopke, “The evolution of chemometrics,” *Anal. Chim. Acta*, vol. 500, no. 1–2, pp. 365–377, Dec. 2003.

[67] J. Luypaert, D. L. Massart, and Y. Vander Heyden, “Near-infrared spectroscopy applications in pharmaceutical analysis,” *Talanta*, vol. 72, no. 3, pp. 865–883, 2007.

[68] G. Reich, “Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications,” *Adv. Drug Deliv. Rev.*, vol. 57, no. 8, pp. 1109–1143, Jun. 2005.

[69] M. Blanco and M. A. Romero, “Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation,” *Analyst*, vol. 126, pp. 2212–2217, 2001.

[70] O. Y. Rodionova, Y. V Sokovikov, and A. L. Pomerantsev, “Quality control of packed raw materials in pharmaceutical industry.,” *Anal. Chim. Acta*, vol. 642, no. 1–2, pp. 222–7, May 2009.

[71] A. Y. Miró Vera and M. Alcalà Bernàrdez, “Near-Infrared Spectroscopy in Identification of Pharmaceutical Raw Materials,” *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, pp. 1–19, 2017.

[72] EMA, “Guideline on the use of Near Infrared Spectroscopy (NIRS) by the pharmaceutical industry and the data requirements for new submissions and variations,” *European Medicine Agency*, vol. 44, no. January. pp. 1–28, 2014.

[73] “5.21. Chemometric methods applied to analytical chemistry,” in *European Pharmacopoeia 8.7*, 2016, pp. 5641–5658.

[74] ICH, “Pharmaceutical Development Q8,” *ICH Harmon. Tripart. Guidel.*, vol. 8, no. August, pp. 1–28, 2009.

[75] M. Fonteyne *et al.*, “Influence of raw material properties upon critical quality attributes of continuously produced granules and tablets,” *Eur. J. Pharm. Biopharm.*, vol. 87, no. 2, pp. 252–263, 2014.

[76] U.S. Department of Health and Human Services (FDA), *Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. 2004, pp. 301–827.

[77] Q. Guo, L. Nie, L. Li, and H. Zang, “Estimation of the critical quality attributes for hydroxypropyl methylcellulose with near-infrared spectroscopy and chemometrics,” *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, vol. 177, pp. 158–163, Apr. 2017.

[78] M. Blanco, R. Cueva-Mestanza, and A. Peguero, “Controlling individual steps in the production process of paracetamol tablets by use of NIR spectroscopy,” *J. Pharm. Biomed. Anal.*, vol. 51, no. 4, pp. 797–804, Mar. 2010.

[79] R. J. Romañach, A. D. Román-Ospino, and M. Alcalà, “A Procedure for Developing Quantitative Near Infrared (NIR) Methods for Pharmaceutical Products,” in *Process Simulation and Data Modeling in Solid Oral Drug Development and Manufacture. Methods in Pharmacology and Toxicology,* M. Ierapetritou and R. Ramachandran, Eds. New York: Humana Press, New York, NY, 2016, pp. 133–158.

[80] P. H. E. Ziémons, J. Mantanus, P. Lebrun, E. Rozet, B. Evrard, “Acetaminophen determination in low-dose pharmaceutical syrup by NIR spectroscopy,” *J. Pharm. Biomed. Anal.*, vol. 53, pp. 510–516, 2010.

[81] M. Blanco and M. Romero, “Near infrared transflectance spectroscopy: Determination of dexketoprofen in a hydrogel,” *J. Pharm. Biomed. Anal.*, vol. 30, no. 3, pp. 467–472, Oct. 2002.

[82] M. Blanco, M. Alcalá, and M. Bautista, “Pharmaceutical gel analysis by NIR spectroscopy. Determination of the active principle and low concentration of preservatives,” *Eur. J. Pharm. Sci.*, vol. 33, no. 4–5, pp. 409–414, 2008.

[83] J. G. Rosas, M. Blanco, J. M. González, and M. Alcalá, “Quality by design approach of a pharmaceutical gel manufacturing process, part 1: Determination of the design space,” *J. Pharm. Sci.*, vol. 100, no. 10, pp. 4432–4441, Oct. 2011.

[84] J. G. Rosas, M. Blanco, J. M. González, and M. Alcalá, “Quality by design approach of a

pharmaceutical gel manufacturing process, part 2: Near infrared monitoring of composition and physical parameters," *J. Pharm. Sci.*, vol. 100, no. 10, pp. 4442–4451, Oct. 2011.

[85] S. Kawata, "Instrumentation for Near-Infrared Spectroscopy," in *Near-Infrared spectroscopy: Principles, Instruments, Applications*, H.W. Siesler, Y. Ozaki, S. Kawata, and H.M. Heise, Eds. Weinheim: Wiley-VCH, 2002, pp. 43–74.

[86] Z. Shi, R. P. Cogdill, S. M. Short, and C. A. Anderson, "Process characterization of powder blending by near-infrared spectroscopy: Blend end-points and beyond," *J. Pharm. Biomed. Anal. J. Pharm. Biomed.*, vol. 47, pp. 738–745, 2008.

[87] T. De Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J. P. Remon, and C. Vervaet, "Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes," *Int. J. Pharm.*, vol. 417, no. 1–2, pp. 32–47, Sep. 2011.

[88] M. Llusá, K. Pingali, and F. J. Muzzio, "Method to study the effect of blend flowability on the homogeneity of acetaminophen," *Drug Dev. Ind. Pharm.*, vol. 39, no. 2, pp. 252–258, Feb. 2013.

[89] L. X. Liu, I. Marziano, A. C. Bentham, J. D. Litster, E.T.White, and T. Howes, "Effect of particle properties on the flowability of ibuprofen powders," *Int. J. Pharm.*, vol. 362, no. 1–2, pp. 109–117, Oct. 2008.

[90] M. Blanco and A. Villar, "Development and Validation of a Method for the Polymorphic Analysis of Pharmaceutical Preparations Using near Infrared Spectroscopy," *J. Pharm. Sci.*, vol. 92, no. 4, pp. 823–830, Apr. 2003.

[91] Á. Gombás, I. Antal, P. Szabó-Révész, S. Marton, and I. Erõs, "Quantitative determination of crystallinity of alpha-lactose monohydrate by Near Infrared Spectroscopy (NIRS)," *Int. J. Pharm.*, vol. 256, no. 1–2, pp. 25–32, 2003.

[92] M. Blanco, D. Valdés, I. Llorente, and M. Bayod, "Application of NIR Spectroscopy in Polymorphic Analysis: Study of Pseudo-Polymorphs Stability," *J. Pharm. Sci.*, vol. 94, no. 6, pp. 1336–1342, Jun. 2005.

[93] T. Koide *et al.*, "Detection of component segregation in granules manufactured by high shear granulation with over-granulation conditions using near-infrared chemical imaging," *Int. J. Pharm.*, vol. 441, no. 1–2, pp. 135–145, 2013.

[94] A. Burggraeve *et al.*, "Development of a fluid bed granulation process control strategy based on real-time process and product measurements," *Talanta*, vol. 100, pp. 293–302, 2012.

[95] P. Frake, D. Greenhalgh, S. M. Grierson, J. M. Hempenstall, and D. R. Rudd, "Process control and end-point determination of a fluid bed granulation by application of near infra-red spectroscopy," *Int. J. Pharm.*, vol. 151, no. 1, pp. 75–80, May 1997.

[96] J. Rantanen, E. Räsänen, O. Antikainen, J.-P. Mannermaa, and J. Yliruusi, "In-line moisture measurement during granulation with a four-wavelength near-infrared sensor: an evaluation of process-related variables and a development of non-linear calibration model," *Chemom. Intell. Lab. Syst.*, vol. 56, no. 1, pp. 51–58, Apr. 2001.

[97] M. Blanco and M. Alcalà, "Simultaneous quantitation of five active principles in a pharmaceutical preparation: Development and validation of a near infrared spectroscopic method," *Eur. J. Pharm.*

Sci., vol. 27, pp. 280–286, 2006.

[98] M. Alcalà, J. Ropero, R. Vázquez, and R. J. Romañach, “Deconvolution of Chemical Physical Information from Intact Tablets NIR Spectra: Two-Three-Way Multivariate Calibration Strategies for Drug Quantitation,” *J. Pharm. Sci.*, vol. 98, no. 8, pp. 2747–2758, Aug. 2009.

[99] M. Blanco, M. Alcalá, J. M. González, and E. Torras, “A process analytical technology approach based on near infrared spectroscopy: Tablet hardness, content uniformity, and dissolution test measurements of intact tablets,” *J. Pharm. Sci.*, vol. 95, no. 10, pp. 2137–2144, Oct. 2006.

[100] M. Blanco and M. Alcalá, “Content uniformity and tablet hardness testing of intact pharmaceutical tablets by near infrared spectroscopy,” *Anal. Chim. Acta*, vol. 557, no. 1–2, pp. 353–359, Jan. 2006.

[101] M. Alcalà, J. León, J. Ropero, M. Blanco, and R. J. Romañach, “Analysis of low content drug tablets by transmission near infrared spectroscopy: Selection of calibration ranges according to multivariate detection and quantitation limits of PLS models,” *J. Pharm. Sci.*, vol. 97, no. 12, pp. 5318–5327, Dec. 2008.

[102] M. Blanco, M. A. Romero, and M. Alcalà, “Strategies for constructing the calibration set for a near infrared spectroscopic quantitation method.,” *Talanta*, vol. 64, no. 3, pp. 597–602, Oct. 2004.

[103] M. Blanco, M. Bautista, and M. Alcalà, “API Determination by NIR Spectroscopy Across Pharmaceutical Production Process,” *AAPS PharmSciTech*, vol. 9, no. 4, pp. 1130–1135, Dec. 2008.

[104] M. Blanco and A. Peguero, “Influence of physical factors on the accuracy of calibration models for NIR spectroscopy,” *J. Pharm. Biomed. Anal.*, vol. 52, no. 1, pp. 59–65, 2010.

[105] L. S. A. Pereira, M. F. Carneiro, B. G. Botelho, and M. M. Sena, “Calibration transfer from powder mixtures to intact tablets: A new use in pharmaceutical analysis for a known tool,” *Talanta*, vol. 147, pp. 351–357, Jan. 2016.

[106] M. Blanco and A. Peguero, “Influence of physical factors on the accuracy of calibration models for NIR spectroscopy,” *J. Pharm. Biomed. Anal.*, vol. 52, no. 1, pp. 59–65, May 2010.

[107] V. Càrdenas, M. Blanco, and M. Alcalà, “Strategies for Selecting the Calibration Set in Pharmaceutical Near Infrared Spectroscopy Analysis. A Comparative Study,” *J. Pharm. Innov.*, vol. 9, no. 4, pp. 272–281, 2014.

[108] V. Cárdenas, M. Cordobés, M. Blanco, and M. Alcalà, “Strategy for design NIR calibration sets based on process spectrum and model space : An innovative approach for process analytical technology,” *J. Pharm. Biomed. Anal.*, vol. 114, pp. 28–33, 2015.

[109] T. Peng *et al.*, “Study progression in application of process analytical technologies on film coating,” *Asian J. Pharm. Sci.*, vol. 10, no. 3, pp. 176–185, 2014.

[110] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, and D. Serrano, “Near-infrared analytical control of pharmaceuticals. A single calibration model from mixed phase to coated tablets.,” *Analyst*, vol. 123, no. 11, pp. 2307–12, Nov. 1998.

[111] C. V. Möltgen, T. Herdling, and G. Reich, “A novel multivariate approach using science-based calibration for direct coating thickness determination in real-time NIR process monitoring,” *Eur. J. Pharm. Biopharm.*, vol. 85, no. 3 PART B, pp. 1056–1063, 2013.

[112] C.-V. Möltgen, T. Puchert, J. C. Menezes, D. Lochmann, and G. Reich, "A novel in-line NIR spectroscopy application for the monitoring of tablet film coating in an industrial scale process," *Talanta*, vol. 92, pp. 26–37, Apr. 2012.

[113] A. L. Pomerantsev, O. Y. Rodionova, M. Melichar, A. J. Wigmore, and A. Bogomolov, "In-line prediction of drug release profiles for pH-sensitive coated pellets," *Analyst*, vol. 136, no. 22, p. 4830, Oct. 2011.

[114] A. Palou, J. Cruz, M. Blanco, J. Tom As, J. De Los Rıós, and M. Alcal, "Determination of drug, excipients and coating distribution in pharmaceutical tablets using NIR-CI," 2012.

[115] 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.

[116] J. Workman and L. Weyer, *Practical guide to interpretive near-infrared spectroscopy*. Taylor & Francis, 2008.

[117] H. Grohganz, D. Gildemyn, E. Skibsted, J. M. Flink, and J. Rantanen, "Towards a robust water content determination of freeze-dried samples by near-infrared spectroscopy," *Anal. Chim. Acta*, vol. 676, no. 1–2, pp. 34–40, Aug. 2010.

[118] C. R. Muzzio, N. G. Dini, and L. D. Simionato, "Determination of moisture content in lyophilized mannitol through intact glass vials using NIR micro-spectrometers," *Brazilian J. Pharm. Sci.*, vol. 47, no. 2, pp. 289–297, Jun. 2011.

[119] H. Grohganz, M. Fonteyne, E. Skibsted, T. Falck, B. Palmqvist, and J. Rantanen, "Role of excipients in the quantification of water in lyophilised mixtures using NIR spectroscopy," *J. Pharm. Biomed. Anal.*, vol. 49, no. 4, pp. 901–907, May 2009.

[120] Y. Li, Q. Fan, S. Liu, and L. Wang, "Simultaneous analysis of moisture, active component and cake structure of lyophilized powder for injection with diffuse reflectance FT-NIR chemometrics," *J. Pharm. Biomed. Anal.*, vol. 55, no. 1, pp. 216–219, Apr. 2011.

[121] J. G. Rosas *et al.*, "NIR spectroscopy for the in-line monitoring of a multicomponent formulation during the entire freeze-drying process," *J. Pharm. Biomed. Anal.*, vol. 97, pp. 39–46, Aug. 2014.

[122] V. del Río, M. P. Callao, M. S. Larrechi, L. M. de Espinosa, J. C. Ronda, and V. Cádiz, "Chemometric resolution of NIR spectra data of a model aza-Michael reaction with a combination of local rank exploratory analysis and multivariate curve resolution-alternating least squares (MCR-ALS) method," *Anal. Chim. Acta*, vol. 642, no. 1–2, pp. 148–154, May 2009.

[123] L. Vann and J. Sheppard, "Use of near-infrared spectroscopy (NIRs) in the biopharmaceutical industry for real-time determination of critical process parameters and integration of advanced feedback control strategies using MIDUS control," *J. Ind. Microbiol. Biotechnol.*, vol. 44, no. 12, pp. 1589–1603, Dec. 2017.

[124] T. Suo *et al.*, "Combining near infrared spectroscopy with predictive model and expertise to monitor herb extraction processes," *J. Pharm. Biomed. Anal.*, vol. 148, pp. 214–223, Jan. 2018.

2. Performance of the Process Spectrum calculated at extreme API concentration values in the inclusion of the process variability of pharmaceutical solids manufacturing into calibration sets of samples prepared at the laboratory

As has been described in section 1.7.1, applications of NIRS in pharmaceutics can be found along the whole industry. In this chapter two examples of such applications are denoted in depth. The first one is a contribution that describes the criteria and approaches involved in the construction of classification models for identification of pharmaceutical raw materials. This essay considers the current recommendations of the EMA and Ph. Eur. regarding this kind of methods, which are illustrated at each step of the process of development of the model, by means of an example based on real samples. This section of the thesis is an example of qualitative applications of NIRS and, even when it does no consist of any innovative results, it provides a useful and updated compilation about the topic. This example has been published during the development of this thesis [1], and can be seen in Annex 1.

The second example is a contribution focused on the performance of an algebraic algorithm employed in the construction of calibration sets for quantitative models developed for pharmaceutical solids. This work states the capability of the *Process Spectrum* (PS) at extreme concentration values (± 30% of the API nominal value) for the inclusion of the process variability into calibration sets created with samples prepared at the laboratory, as will be exposed below.

2.1 Introduction

Accuracy and robustness are the assessment criteria most resolutely sought during the construction of quantitative models based on NIRS data. These criteria are particularly relevant if such methods are intended to be used as a process analytical technology (PAT) tool in the pharmaceutical industry. Reaching accurate and robust calibrations is highly dependent on the data incorporated in the calibration sets during the development of the model. The preparation of calibration sets with enough variability can comprise, in many cases, the need of a large amount of data with a large variation, something generally difficult to obtain in real situations [2]. In fact, this requirement has been indicated as one of the main disadvantages of applying NIRS to pharmaceutical process control [2].

Because of that, many efforts have been done focus on including as much of the pharmaceutical process variability into the calibration sets of models based on NIRS. A summary of the literature available about this issue from 1987 to 2009 is presented in [3]. Since the publication of regulatory considerations on the use of NIRS technique as a PAT tool, promoted by the US Food and Drug Administration (FDA) in 2004, and the European Medicine Agency (EMA) in 2014, this aspect has been even studied each time with more attention. Among the most relevant contributions of 2004, it is possible to find those oriented to the use of calibration sets based on samples prepared at the laboratory, with expanded concentration range of the analytes of interest, intending to overcome the difficulties stated by the narrow range of chemical variability generated by samples directly taken from production lines [4]–[6]. However, this strategy by itself did not include any information about the physical changes occurred during the process, such as granulation, compaction or coating, aspects with proved influence on the accuracy of NIRS calibration models [7], [8]. As an alternative to this condition, Blanco et al [4], proposed the preparation of mixed calibration sets, joining laboratory and production samples with the objective of involving a wider concentration range at the same time than physical changes intrinsic to the process [5]. This strategy stands an alternative to the generation of samples in pilot plant facilities, which is another proved useful path [9] but also a very expensive one.

Even when these methodologies improved with statistical significance the accuracy of the resulting models, the preparation of samples for calibration implied the need of laboratory work for the preparation of powder samples, besides over and under dosing of production samples. This was one of the motivations for developing the *Process Spectrum* (PS) tool in 2010. The PS allows the incorporation of the physical variability of a pharmaceutical process by means of a mathematical algorithm [3]. This strategy includes the spectral differences between laboratory and production samples as a vector that can be added to powder samples spectra in proportions that can be established based on a multiplicative factor *m* [10]. In 2014, this strategy was compared with advantaged results over the mixed calibration sets of laboratory and production samples [11]. Furthermore, in 2015, additional studies were conducted on the selection of the mentioned multiplicative factor *m*, both for a better understanding of its effect as for the description of criteria for selecting it [12]. All these previous efforts have demonstrated the capability of the PS strategy for including the process variability of pharmaceutical solids manufacturing into calibration sets based on samples prepared at the laboratory. In all these contributions the PS have been calculated only at the API nominal concentration value. The purpose of the present work is to study the effect of calculating the PS at extreme points of a concentration range of ±30 % relative to the nominal API concentration value, and to evaluate its advantages and disadvantages in the optimization of quantitative PLS models.

2.2 Experimental

2.2.1 Samples preparation

A total of 350 laboratory samples distributed in five concentration levels along the API nominal range 7 - 13% w/w were prepared (±30 regarding the nominal value, 10%w/w). For this purpose, seven placebo mixtures were set combining lactose (67.5%w/w), microcrystalline cellulose (30%w/w), povidone (2%w/w) and magnesium stearate (constant at 0.5%w/w). A full factorial design was employed for minimizing collinearity between excipients concentrations as much as possible, while spanning their concentrations in ±5% (see Table 1). A mechanical shaker was employed to ensure a homogeneous blend of the excipients.

Granulated samples were prepared by spreading 5%w/w of water on separated portions of the powder samples, using manual stirring followed by oven dry at 40ºC and 0.1bar, during 24 h. The tablets were prepared both from powder and from granulated samples. All of them were compressed at 100Mpa using a Perkin-Elmer press. The total amount of 350 samples is the result of 35 powder samples, 35 granulated samples, 140 tablets prepared from powder and 140 tablets prepared form granulated (4 tablets replicates were made per sample with the aim of including the variability due to compaction pressure).

Table 1. Correlation factors of concentrations between the components of the samples

	API	**Lactose**	**MCC**	**Povidone**
API	1			
Lactose	-0.4	1		
Microcrystalline cellulose	-0.3	-0.8	1	
Povidone	-0.4	0.2	0.08	1

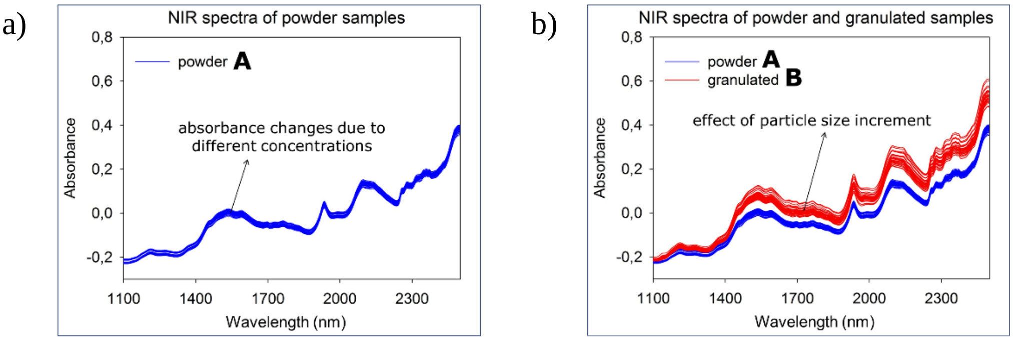
2.2.2 NIR spectra acquisition and chemometrics software

A FT-NIR spectrometer by FOSS NIRSystems, Inc., model 5000, governed by the software Vision 2.51 (Denmark), was used for the spectral data acquisition. Each spectrum is the result of an average of 32 scans in the range from 1100 to 2500 nm, with wavelength intervals of 2 nm. For powder and granulated, three replicates were acquired per sample, with manual stirring between them for guarantying representativeness of the whole mixture from the irradiated particles. For tablets, one spectrum of each face was acquired. The exploratory analysis, the selection of the calibration sets and the calculation of the PLS models was completed using Unscrambler 10.3 from CAMO (Norway).

2.3 Results and discussion

2.3.1 Physical differences between the samples expressed in the NIR absorbance spectra

Besides the variations in the intensities of the absorption bands due to chemical changes promoted by the diverse compositions of the samples, the resultant spectra of the four types of the samples exhibited clear evidences of their physical differences. Figure 1 shows all the absorbance spectra, naming spectra of powder samples with letter **A**, spectra of granulated samples with letter **B**, tablets from powder samples with the letter **C** and tablets from granulated with a **D**. Granulated samples generated spectra with less scattering effect regarding the powder samples (in consequence higher registered absorbance), because of the increment of the particle size. The compression effect was also exposed by a baseline shift between the tablets and the samples of particulates.



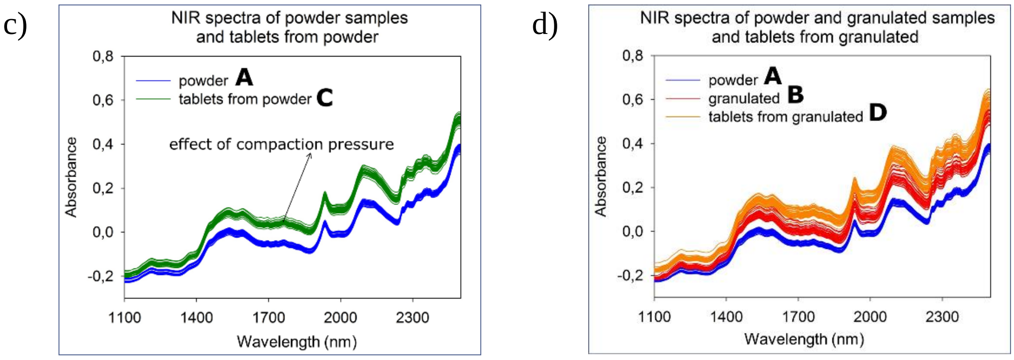


Figure 1. NIR absorbance spectra of a) powder samples; b) powder and granulated samples; c) powder samples and tablets from powder and d) powder and granulated samples and tables from granulated.

The PS vectors were calculated following the equation detailed in [3], PS=S_a-S_b, where S_a=spectra after physical change and S_b=Spectra before physical change. Figure 2a) presents the plots of the PS obtained from averaged spectra at extreme and centre concentration levels of the API concentration range (7, 10 and 13%w/w), using the absorbance spectra of the diverse physical changes. PS of granulation effect was calculated by subtraction of powder samples spectra from the granulated samples spectra (B-A). The PS of compression effect was calculated for the two kind of original samples of particulates, i.e. powder and granulated. C-A represents the subtraction of spectra of powder samples from tablets prepared from powder samples, and D-B granulated samples spectra from spectra of tablets from granulated samples. It is possible to observe several differences in the resulting PS at the extreme concentration levels regarding the one obtained at the central nominal value.

Based on the proved capabilities of classical spectral data pre-treatments, such as Standard Normal Variate (SNV), for correcting spectral differences due to scattering effects, it was tried to calculate the PS also after using this pre-treatment. Figure 2b) exposes that, at extreme API concentration values, not all the PS differences can be corrected by SNV.

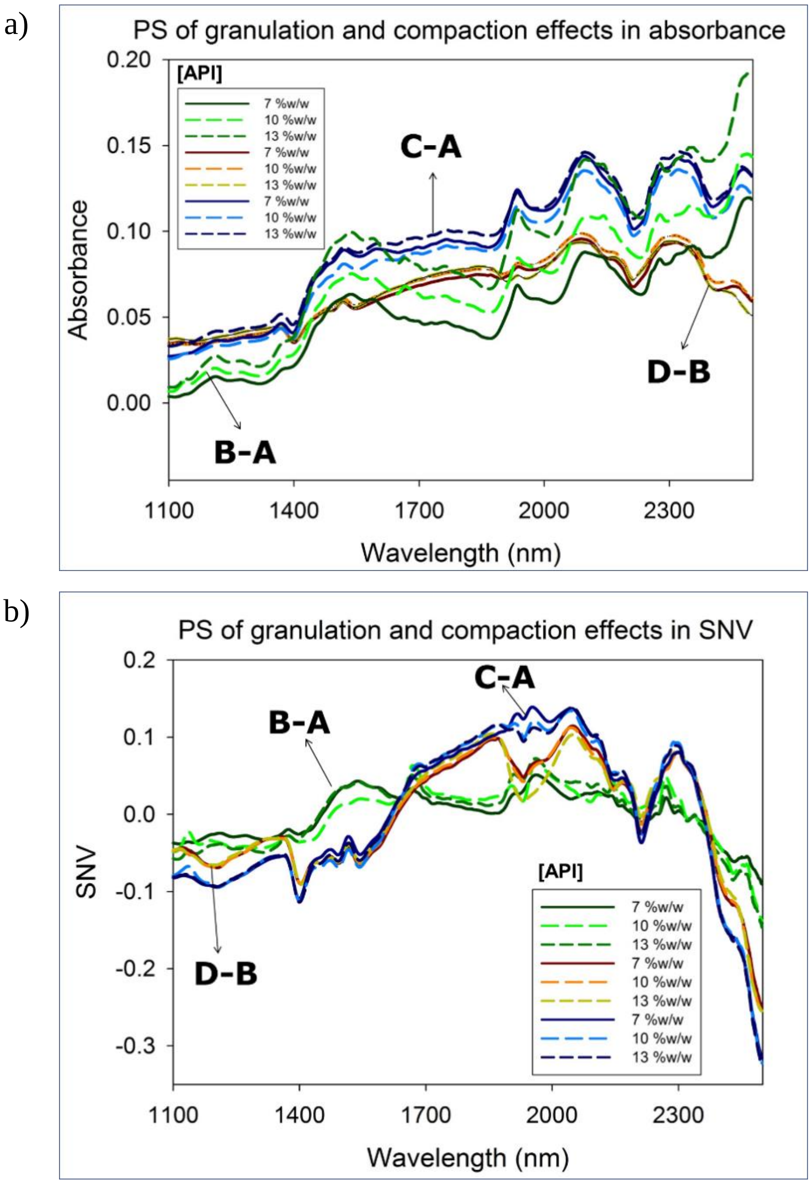


Figure 2. Process spectrum calculated at extreme concentrations points of the studied range, contrasted with the nominal concentration value (10%w/w) a) absorbance; b) after SNV pre-treatment.

A more detailed perspective of the effect of the API concentration in the PS, can be observed in Figures 3 and 4. Figure 3 illustrates the changes in absorbance generated in the NIR spectra of powder samples when the PS of granulation is added at three different API concentration values (7, 10 and 13%w/w). The PS obtained from compacted samples are presented in Figure 4. The differences between the spectra modified by the addition of the PS can be observed in the slope of the spectral profile of the set of powder samples, which is incremented directly proportional to the increment of the API concentration.

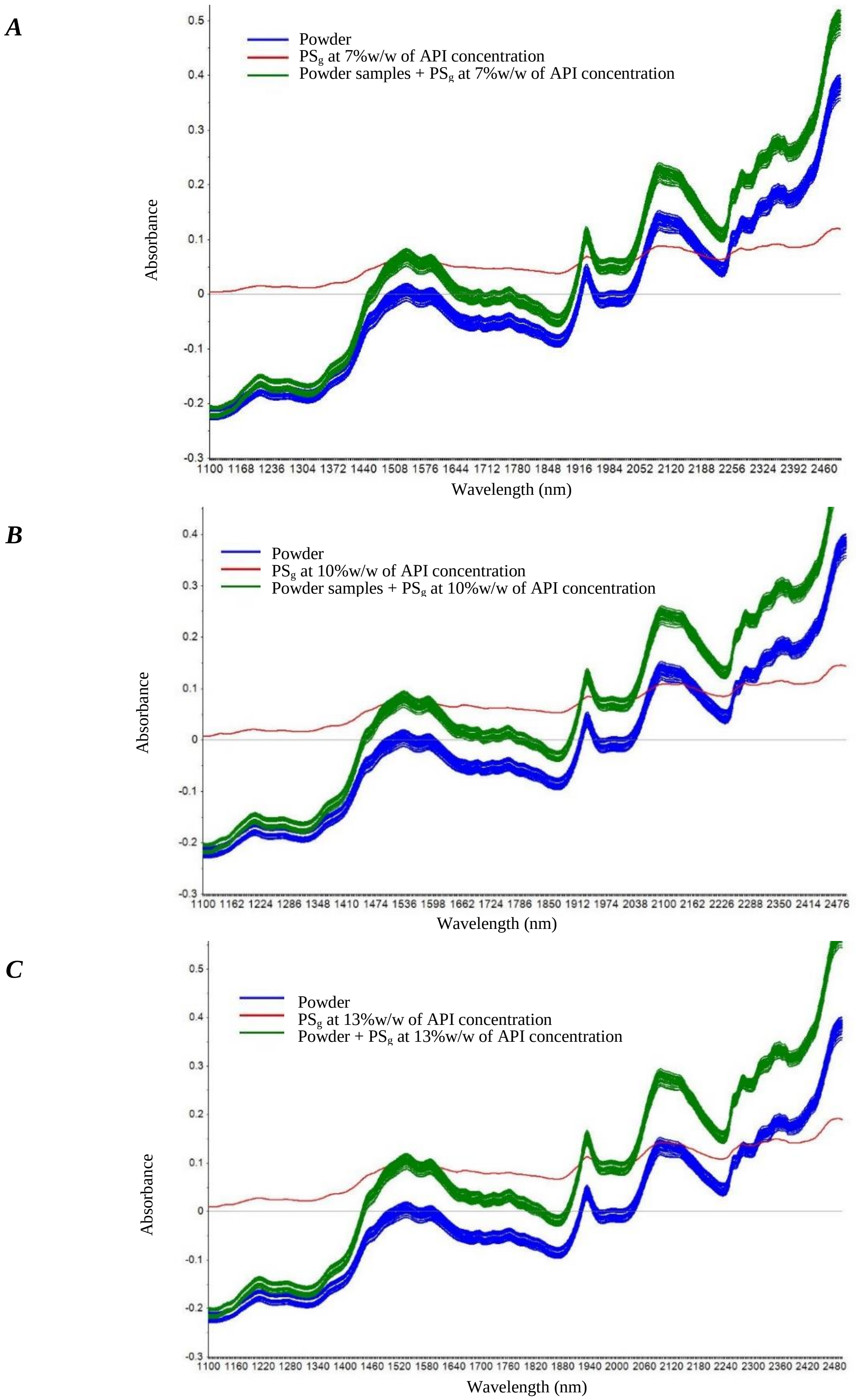


Figure 3. Absorbance NIR spectra of powder samples with and without PS of granulation process (PS_g), calculated at the API concentrations of: 7%w/w (***A***); 10%w/w (***B***); 13%w/w (***C***).

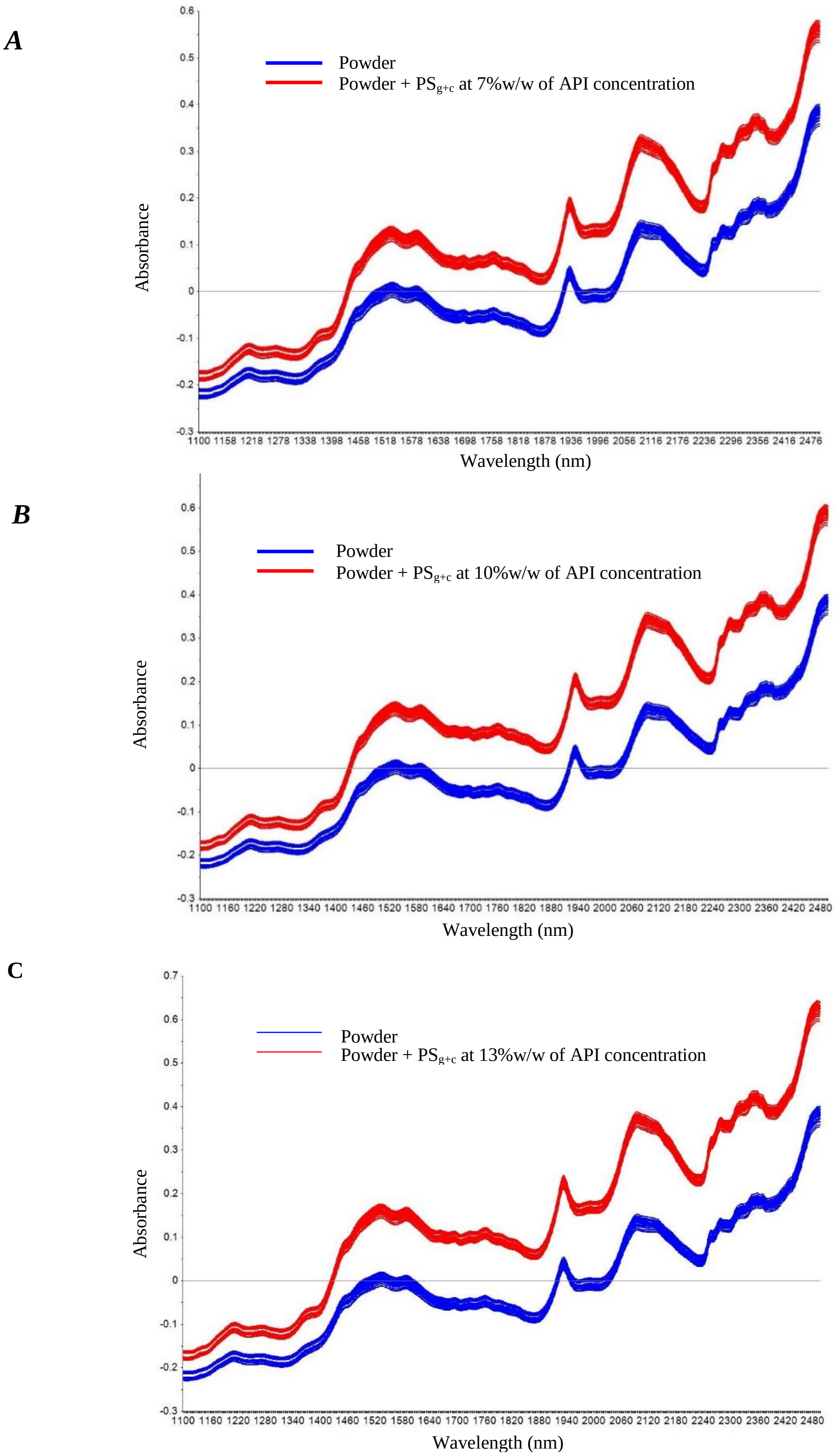


Figure 4. Absorbance NIR spectra of powder samples and modified samples with PS of granulation and compaction processes (PSg+c), calculated at the API concentrations of 7%w/w (***A***); 10%w/w ***(B***); 13%w/w (***C***).

The differences related to the granulation process -increment on the slope of the spectral profiles observed in Figure 3- and the additional shift of the whole baseline produced by the compaction process illustrated in Figure 4, can be also noticed in Figure 5. This figure shows all the PS in absorbance, calculated at five API concentration levels over the whole spectral range studied. This figure shows that the concentration of the API provides an additional source of variability to the PS, besides the physical changes devoted to the manufacturing process which is intended to be represented.

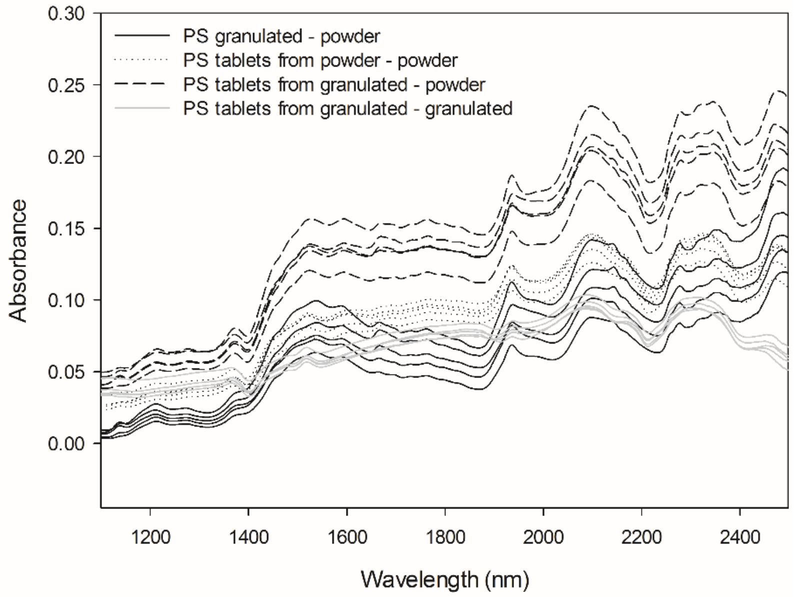


Figure 5. Scores plot of PCA over the whole spectral range (1100-2498 nm) of PS calculated in absorbance for all the physical changes studied. Number of samples displayed are organized regarding the increment of API concentration (7, 8.5, 10, 11.5 and 13%w/w) inside of each group of samples.

A more detailed view of these results can be found in the scores plot of a PCA of the PS calculated, including the distinction between the five levels of API concentration studied, as can be seen in Figure 6. This plot exposes the particle size and the compaction pressure as the main variability sources for the first two principal components. However, it is also possible to observe that the first component (PC1, explaining the 94% of the variability) is related to the API concentration.

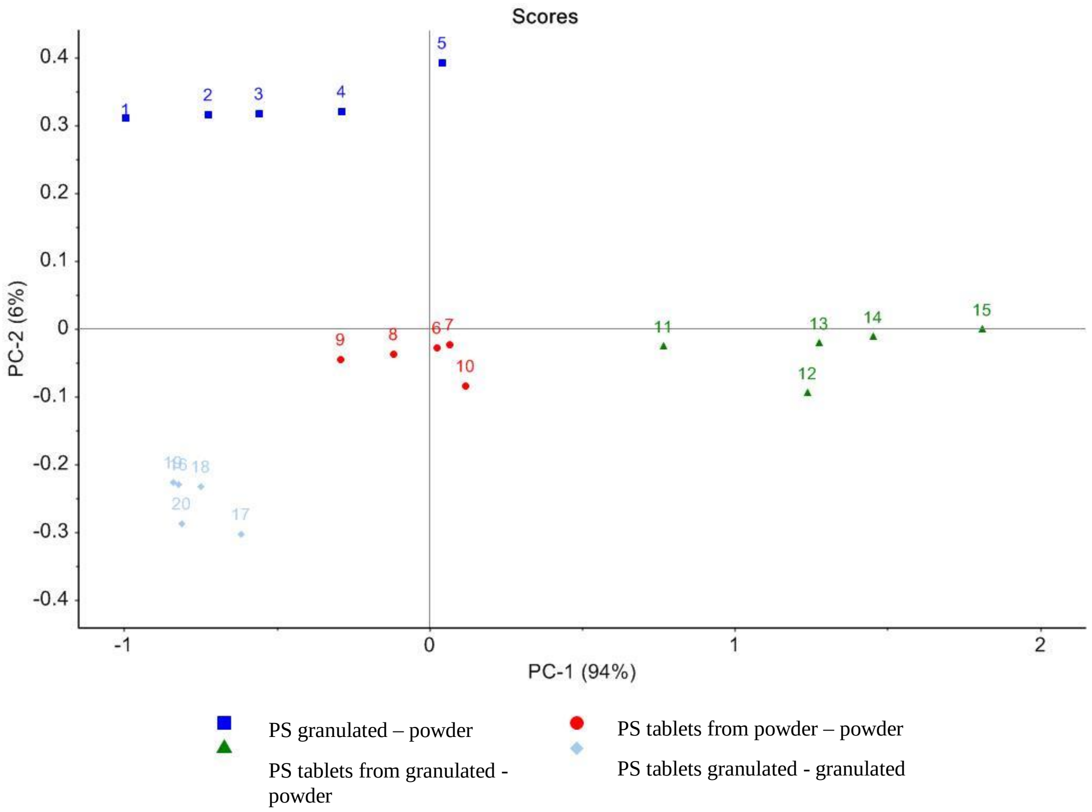


Figure 6. Scores plot of PCA over the whole spectral range (1100-2498 nm) of PS calculated in absorbance for all the physical changes studied. Number of samples displayed are organized regarding the increment of API concentration (7, 8.5, 10, 11.5 and 13%w/w) inside of each group of samples.

2.3.2 Evaluation of predictive capability of PLS models

The effect of the API concentration in the PS was also studied regarding the predictive capability of PLS models calculated using the powder samples, and subsequently modified with the PS at diverse API concentration values. Table 2, shows the errors of predictions and the media of residuals of a PLS model constructed using only powder samples, in the prediction of the granulated, tablets compacted from powder and tablets compacted from granulated. Figure 7, shows the graphic of Explained variance per number of factors of this model, obtained using the Gap Derivative (GD) pre-treatment of second order, segment size of 1 and the spectral range 1616-2180 nm. It is worth mentioning that these model parameters were found after trying the application of classical data pre-treatments (SNV, Savitzky-Golay derivatives among others) as well as many diverse combinations of them.

Additionally, Table 2 presents the results obtained from the same model modified by the inclusion in the calibration set of the PS at diverse API concentration levels. All the models were created using the leave one out strategy of cross-validation (CV), and in all the predictions presented in Table 2, 7 PLS factors were employed.

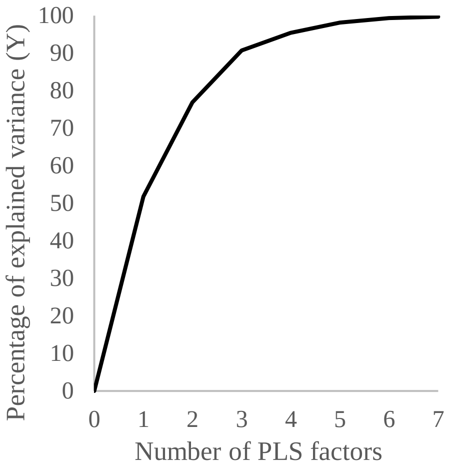


Figure 7. Explained variance per number of PLS factors of the model created using the powder samples as calibration set, GD pre-treatment of second order and segment size of 1, in the spectral range 1616-2180 nm and using CV.

Table 2. Root mean square error of calibration (RMSEC), prediction (RMSEP) and media of residuals for the prediction of granulated and tablets from granulated samples using the calibration set of powder samples with and without application of the PS strategy, at different API concentration values.

Calibration set	RMSEC (%w/w)	Samples predicted	RMSEP (%w/w)	$\overline{X}$ of residuals
Powder	0.17	Powder	0.02	0.00
		Granulated	0.53	-5.0
		Tablets from powder	0.22	-1.8
		Tablets from granulated	0.48	-4.7
Powder + PS$_g$ at 7%w/w of API	0.14	Granulated	0.16	-0.09
Powder + PS$_g$ at 10%w/w of API	0.16	Granulated	0.16	0.2
Powder + PS$_g$ at 13%w/w of API	0.17	Granulated	0.19	1.2
Powder + PS$_{g+c}$ at 7%w/w of API	0.01	Tablets from granulated	0.13	-0.1
Powder + PS$_{g+c}$ at 10%w/w of API	0.02	Tablets from granulated	0.14	-0.5
Powder + PS$_{g+c}$ at 13%w/w of API	0.02	Tablets from granulated	0.19	1.4

The changes in the trend of residuals obtained from the prediction of the four kind of samples when they were predicted using the models indicated in Table 2, are illustrated by the plots presented in Figure 8. This figure clearly shows the effect of the API concentration at which each PS was calculated.

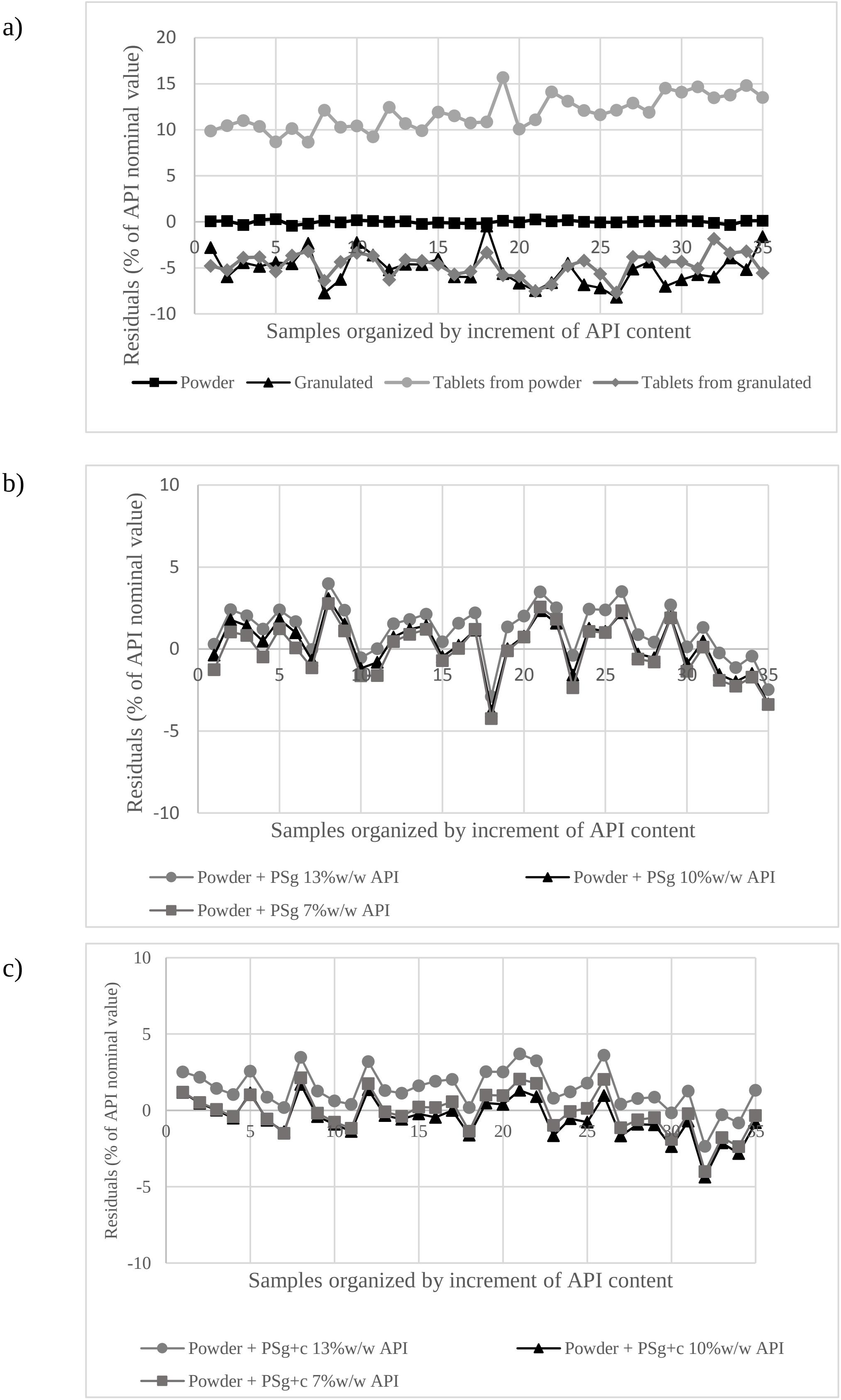


Figure 8. Evolution of residuals in predictions using models created using a) powder samples without correction of PS; b) powder samples with PS_g at 7, 10 and 13%w/w of API; c) powder samples with PS_{g+c} at 7, 10 and 13%w/w of API

In Figure 8, the samples are displayed covering the API contents of 7, 8.5, 10, 11.5 and 13%w/w. Figures 8b and 8c, show the effect of the correction provided by the PS changes over the API concentration range. The process variability due to the granulation, included by PS_g, produced similar residuals when the PS_g was calculated at 7 and 10%w/w API concentration values, regarding the values obtained when the PS_g was calculated at 13%w/w. This trend can be also observed when the PS was calculated after the compaction process.

2.4 Conclusions

Due to the NIR spectra can be affected not only by physical, but also by chemical changes during the pharmaceutical manufacturing process, the performance of the PS strategy is affected by changes in the concentration of the components of solid blends. Based on the presented results, it is possible to confirm that the calculation of the PS at the nominal API concentration is the most appropriate strategy for quantitative modelling purposes. Additionally, these results suggest the possibility of studying the PS as a pre-treatment potentially useful for modelling based on samples prepared at the laboratory. A macro for the calculation of the simple subtraction of the PS algorithm would be suitable for applying to new samples corrections due to the process variability in an fast and reliable way.

2.5 References

[1] A. Y. Miró Vera and M. Alcalà Bernàrdez, “Near-Infrared Spectroscopy in Identification of Pharmaceutical Raw Materials,” *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, pp. 1–19, 2017.

[2] M. Jamrógiewicz, “Application of the near-infrared spectroscopy in the pharmaceutical technology,” *J. Pharm. Biomed. Anal.*, vol. 66, pp. 1–10, 2012.

[3] M. Blanco and A. Peguero, “Analysis of pharmaceuticals by NIR spectroscopy without a reference method,” *TrAC Trends Anal. Chem.*, vol. 29, no. 10, pp. 1127–1136, Nov. 2010.

[4] M. Blanco, M. A. Romero, and M. Alcalà, “Strategies for constructing the calibration set for a near infrared spectroscopic quantitation method.,” *Talanta*, vol. 64, no. 3, pp. 597–602, Oct. 2004.

[5] M. Blanco, M. Bautista, M. Alcala, and A. C. Unit, “Preparing Calibration Sets for Use in Pharmaceutical Analysis by NIR Spectroscopy,” vol. 97, no. 3, pp. 1236–1245, 2008.

[6] M. Blanco, M. Bautista, and M. Alcalà, “API Determination by NIR Spectroscopy Across Pharmaceutical Production Process,” *AAPS PharmSciTech*, vol. 9, no. 4, p. 1130, 2008.

[7] M. Alcal?, M. Blanco, M. Bautista, and J. M. Gonz?lez, “On-line monitoring of a granulation process by NIR spectroscopy,” *J. Pharm. Sci.*, vol. 99, no. 1, pp. 336–345, Jan. 2010.

[8] M. Blanco and A. Peguero, “Influence of physical factors on the accuracy of calibration models for NIR spectroscopy,” *J. Pharm. Biomed. Anal.*, vol. 52, no. 1, pp. 59–65, 2010.

[9] D. Jouan-Rimbaud, B. Walczak, D. L. Massart, I. R. Last, and K. A. Prebble, “Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data,” *Anal. Chim. Acta*, vol. 304, no. 3, pp. 285–295, 1995.

[10] M. Blanco, R. Cueva-Mestanza, and A. Peguero, “NIR analysis of pharmaceutical samples without reference data: Improving the calibration,” *Talanta*, vol. 85, no. 4, pp. 2218–2225, 2011.

[11] V. Càrdenas, M. Blanco, and M. Alcalà, “Strategies for Selecting the Calibration Set in Pharmaceutical Near Infrared Spectroscopy Analysis. A Comparative Study,” *J. Pharm. Innov.*, vol. 9, no. 4, pp. 272–281, 2014.

[12] V. Cárdenas, M. Cordobés, M. Blanco, and M. Alcalà, “Strategy for design NIR calibration sets based on process spectrum and model space : An innovative approach for process analytical technology,” *J. Pharm. Biomed. Anal.*, vol. 114, pp. 28–33, 2015.

[13] “Transfer of multivariate calibration models: a review,” *Chemom. Intell. Lab. Syst.*, vol. 64, no. 2, pp. 181–192, Nov. 2002.

3. Comparison of the performance of bench-top and hand-held NIRS instruments concerning the identification of new psychoactive substances

3.1 Introduction

According to the World Drug Report 2017, issued by the United Nations Office on Drugs and Crime (UNODC), the current synthetic drugs market is the most complex and widely spread over the recent years. The substances available have expanded considerably, with the persistence of traditional drugs and the emergence of new psychoactive substances (NPS) every year. The main public health risks of this polydrug phenomenon are outlined by the unavailability of scientific information about the NPS -all these compounds are introduced into the market without any published *in vivo* testing even in animal models-, the wide variation in both the quantity and effectiveness of their active components and the potential combinations that can be used. Between 2009 and 2016, 106 countries and territories reported the emergence of 739 different NPS to the UNODC, marketed in many different ways, most of them with lower prices than controlled drugs and easier accessibility (through internet) [1].

One of the factors that have made the control of NPS harder are the molecular differences that can been found between the structures of illicit controlled drugs and these new compounds. NPS encompass a large number of compounds that can describe diverse substance classes as synthetic cannabinoids, phenethylamines, cathinones, tryptamines, and piperazines. The possibilities of modifying the structure and functionalities of already known illicit substances based on such range of molecules are vast [2]. Even when there is a core group of around 80 NPS that have been reported every year during 2009-2015, and seems to become established in the global market, while some others have disappeared, the general trend for NPS has been increasing over the years. Such evolution of these synthetic drugs and the important public health risk that they represent, raises the need for improved forensic capacity and new approaches for data collection and seized samples identification [1].

To answer this relevant need, diverse analytical approaches have been developed in recent years. Some of them have been focused on the deep characterization of specific NPS structures in products sold as bath salts, potpourri, incense and food of plant origin among other presentations. For this purpose, morphological and chemical techniques have been employed, as well as DNA references for diverse plant species and the confirmation of the MS results by comparison with the Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) mass spectral library [3]. Other contributions have pointed to the identification of NPS in seized samples without reference standards using GC/MS, LC/HRMS, NMR [4] and optical techniques as ATR-IR and Raman spectroscopies [5]. The NIRS has been considered to face this issue as well, due to the relevant reduction of time of analysis and sample preparation that it can offer regarding other concentration techniques.

NIRS instruments can generate a *"fingerprint"* spectrum of each seized NPS, as well as characteristic spectra of the pure components employed during NPS manufacturing processes. This fact allows the development of classification models based on diverse pattern recognition methods using NIR spectra of NPS. This kind of approach have been reported in the literature, using both bench-top and more recently, miniaturized NIRS instruments. Some examples of published applications of bench-top instruments in this field are the identification of illicit controlled drugs as heroin in seized samples [6], [7], and amphetamines [8], cellulose and lactose [9] in ecstasy tablets. For seized ecstasy tablets, even the concentration of diverse amphetamines have been quantified by means of data acquired using both diffuse and transmission acquisition modes, with successful validation using new seized samples [10]. Other traditional drugs as cocaine have been also detected in seized samples using NIRS and chemometrics [11]. Additionally, synthetic molecules belonging to the classes of synthetic cannabinoids and phenethylamines have been identified using NIRS and principal components analysis. For this purpose, a bench-top instrument was used for recording in the spectral range 1000 - 2500 nm, and Principal Component Analysis (PCA) allowed the use of the resulting data for determining and distinguishing indole and indazole derivatives in emerging streets drugs matrices. This work was developed considering 22 synthetic cannabinoids and phenethylamines, 9 from pure synthetic standards and 13 from seized samples. In this case, the confirmation of the NIRS identification results was done using GC/MS [12]. More recently, the prediction of the concentration of AKB48 in samples prepared using the standard reference substance mixed with herbs have been also reported. In this case, the use of herbs allowed the inclusion of some of the complexity of the real matrix [13].

The capability of portable NIRS instruments for identification of NPS have been also studied for diverse researchers [14], [15]. Tsujikawa *et al* used a portable instrument -set for the spectral range 1400-2400 nm- for constructing a spectral library involving 120 pure standard drugs and non-psychoactive drugs. This study showed that the effect of the particle size can be handled by means of mathematical pre-treatments. The validation of the final library was done with 1 real seized sample and 10 samples obtained by internet [14]. The second report using portable instruments was completed by Pederson *et al*, who demonstrated the capability of hand-held instruments for identifying NPS with a very low rate of misidentification, as well as the ability of completing successful calibration transfers between different instruments. Additionally, they were able to identify the individual compounds that constituted unknown mixtures of diverse controlled psychotropic substances (such as cocaine, heroin, oxycodone and diazepam) with paracetamol, caffeine and lidocaine, by means of the use of the Net Analyte Signal (NAS) assessment method [15].

Based on the efforts mentioned above, and focused on the demonstration of the strengths of hand-held NIRS instruments for enhancing on-site NPS control actions of police officers, the aim of this work was to compare the identification capability of bench-top and hand-held instruments based on the performance of classification models of NPS. For this purpose, the same set of NPS samples was analysed using both instruments, and spectral libraries were constructed based on each data set. The performance of both instruments is compared in terms of simplicity of the structure of the resulting spectral libraries and their selectivity values.

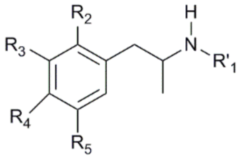
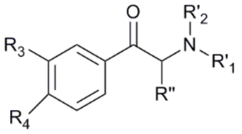
3.2 Experimental

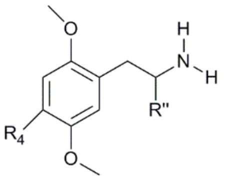
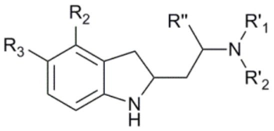
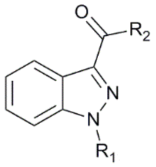
3.2.1 Samples

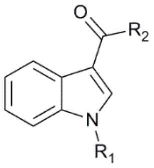
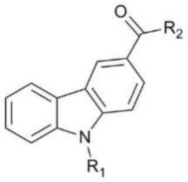
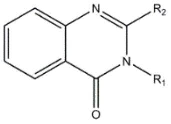
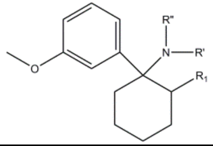
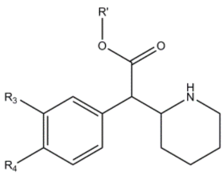
Samples were seized from 2014 to 2017, by the Spanish law enforcement bodies in police raids in Valencia, Spain. The Pharmaceutical Inspection and Drug Control Unit of the Ministry of Home Affairs of Spain provided this sample set to the Department of Analytical Chemistry, University of Valencia, Spain. These samples were kindly provided to the Applied Chemometrics Research Group UAB by the Department of Analytical Chemistry, University of Valencia, Spain. Seized samples consisted of fine white and yellowish powder placed in different plastic bags, most of them with purity over 95%. Classes of the psychoactive substances were amphetamine derivatives (8 compounds), cathinones (17 compounds), 2C-family (4 compounds), tryptamines (6 compounds), synthetic cannabinoids including indazoles, indoles and carbazoles (11 compounds), quinazolinones (2 compounds), arylcyclohexylamines (2 compounds), phenidates (2 compounds), and miscellaneous (3 compounds). Details of employed NPS are provided in Table 1.

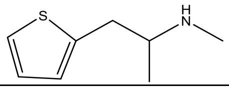
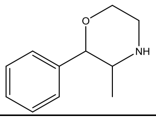
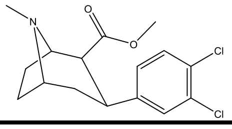
No sample preparation was necessary for NIRS measurements. The samples were transferred to transparent borosilicate glass vials of 32 mm x 11.6 mm closed with plastic tops. The amount of sample varied between 20 and 100 mg, depending on the amount available from each seizure case.

Table 1. Description of the NPS samples analyzed using the bench-top and hand-held NIRS instruments.

SUBSTANCES	STRUCTURE			MOLECULAR WEIGTH (g/mol)
Amphetamine derivatives				
2-FA	R_2=F			153.2
3-FA	R_3=F			153.2
2-FMA	R_1’= methyl	R_2=F		167.2
4-FMA	R_1’= methyl	R_4=F		167.2
6-APB	R_3,R_4=furan			175.2
3-methoxymethamphetamine	R_1’=methyl	R_3=methoxy		179.3
3-FEA	R_1’=ethyl	R_3=F		181.2
5-EAPB	R_1’=ethyl	R_3,R_4=furan		203.3
Cathinone derivatives				
Ethcathinone	R_2’=ethyl	R”=methyl		177.3
3-MMC	R_2’=methyl	R_3=methyl	R”=methyl	177.3
4-MMC (mephedrone)	R_2’=methyl	R_4=methyl	R”=methyl	177.3
3-FMC	R_2’=methyl	R_3=F	R”=methyl	181.2
4-MEC	R_2’=ethyl	R_4=methyl	R”=methyl	191.3
4-MeMABP (4-methylbuphedrone)	R_2’=methyl	R_4=methyl	R”=ethyl	191.3
4-CMC	R_2’=methyl	R_4=Cl	R”=methyl	197.7
Methylone	R_2’=methyl	R_3,R_4=methyl enedioxy	R”=methyl	207.2

Butylone	R_2’=methyl	R_3,R_4=methyl enedioxy	R”=ethyl	221.2
α-PVP	R_1’,R_2’=pyrrolidin	R”=n-propyl		231.3
α-PHP	R_1’,R_2’=pyrrolidin	R”=n-butyl		245.4
MDPPP	R_1’,R_2’=pyrrolidin	R_3,R_4=methyl enedioxy	R”=methyl	247.3
MPHP	R_1’,R_2’=pyrrolidin	R_4=methyl	R”=n-butyl	259.4
MDPV	R_1’,R_2’=pyrrolidin	R_3,R_4=methyl enedioxy	R”= n-propyl	275.3
4-MePPP	R_1’,R_2’=pyrrolidin	R_4=methyl	R”= methyl	
PV9	R_1’,R_2’=pyrrolidin	R”= n-hexyl		309.9
3,4-MDPHP	R_1’,R_2’=pyrrolidin	R_3,R_4=methyl enedioxy	R”= n-butyl	325.8
2C-Family				
2C-E	R_4=ethyl		R”=H	209.3
2C-C	R_4=Cl		R”=H	215.7
2C-P	R_4=n-propyl		R”=H	223.3
DOC	R_4=Cl		R”=methyl	229.7
Tryptamine derivatives				
alpha-Methyltryptamine (AMT)		R”=methyl		174.2
4-OH-MET	R_1’=methyl	R_2’=ethyl	R_2=hydroxy	218.3
5-MeO-DMT	R_1’=methyl	R_2’=methyl	R_3=methoxy	218.3
DPT	R_1’=n-propyl	R_2’=n-propyl		244.4
4-AcO-DMT	R_1’=methyl	R_2’=methyl	R_2=acetoxy	246.3
5-MeO-MIPT	R_1’=methyl	R_2’=iso-propyl	R_3=methoxy	246.3
Synthetic cannabinoids (indazole family)				
THJ-2201	R_1=5F-pentyl		R_2=naphthyl	360.2

CUMYL-4CN-BINACA	R_1=4CN-butyl	R_2=-NH-cumyl		360.2
ADB-CHMINACA	R_1=methylcyclohexyl	R_2=-NH-tert-butyl-carbamoyl		370.2
5F-ADB	R_1=5F-pentyl	R_2=-NH-tert-butyl-methoxycarbonyl		377.2
5F-NPB-22	R_1=5F-pentyl	R_2=-O-quinolinyl		377.2
Synthetic cannabinoids (indole family)				
UR-144	R_1=pentyl	R_2=2,2,3,3-tetramethyl- cyclopropyl		311.4
RCS-4	R_1=pentyl	R_2=4-methoxyphenyl		321.4
JWH-210	R_1=pentyl	R_2=4-ethylnaphthyl		369.5
JWH-081	R_1=pentyl	R_2=4-methoxynaphthyl		371.5
MMB-CHMICA	R_1=methylcyclohexyl	R_2=isopropyl-methoxycarbonyl		370.2
Synthetic cannabinoids (carbazole family)				
MDMB-CHMCZCA	R_1=methylcyclohexyl	R_2=-NH-tert-butyl-methoxycarbonyl		434.3
Quinazolinones				
Etaqualone	R_1=2-ethyl-phenyl		R_2=methyl	264.3
Mebroqualone	R_1=2-bromo-phenyl		R_2=methyl	315.2
Arylcyclohexylamines				
Methoxetamine	R_1=ketone	R’=H	R”=ethyl	247.3
3MeO PCP	R_1=H	R’,R”=cyclohexyl		273.4
Phenidates				
Ethylphenidate	R_3=H	R_4=H	R’=ethyl	247.3
Threo-4-fluoromethylphenidate	R_3=H	R_4=F	R’=methyl	251.3

Miscellaneous		
Methiopropamine		155.3
3-fluorophenmetrazine		195.2
Dichloropane (RTI-111)		328.2

3.2.2 NIRS Instruments

Bench-top instrument: NIR spectrometer from Foss (Denmark), model 5000, fitted with a Rapid Content Analyzer model 6500, employed with diffuse reflectance acquisition mode in the wavelength range of 1100-2498 nm. Each spectrum was the result of 32 scans acquired with a wavelength intervals of 2 nm. The measurements were done in two different days by the same analyst. The software Vision, version 2.51 was used for controlling the instrument and data acquisition.

Hand-held instrument: MicroNIR portable spectrometer from JSDU (USA), model 1700. This device of 45 mm of diameter x 42 mm of height and 60 g of weight, allows the acquisition in reflectance mode along the range 908.1-1676.2 nm, with a spectral sampling interval of 6.25 nm per pixel (detector of 128 pixels). The software employed for controlling the instrument and data acquisition was the IRSE, version 1.3.5.

Software for spectral libraries development: Opus from Bruker, version 7.5 was employed for the construction of both libraries. Previous data visualization and adjustment of files format were done using The Unscrambler from Camo, version 10.3 and PLS Toolbox from Eigenvector, version 8.2.1.

3.2.3 Methodology

a) NIRS measurements

The samples were presented to both NIR spectrometers directly in borosilicate vials, without any physical or chemical pre-treatment. Two replicates were recorded per sample in two different days by the same analyst. Therefore, from each instrument, 4 replicates were available per sample. This data set was divided as follows: those spectra acquired on day one, were employed for preparing the calibration sets and those acquired on day two were employed for validation sets.

The calculation of both spectral libraries was done following the general chemometric strategies intended for describing mathematical algorithms and parameters able to classify new spectra as similar or different from those included in a calibration set [16], [17]. For this purpose, the software OPUS allows the visualization of the original spectra after diverse mathematical pre-treatments, as Standard Normal Variate (SNV), first (1D) and second (2D) Savitzky-Golay derivatives of second polynomial order -with diverse values options for the number of points to be included in the smoothing window calculation-, as well as combinations of them. Based on that, the differences between samples in the whole spectral range available in each case (1100-2498 nm for the bench-top instrument and 908.1-1676.2 nm for the hand-held instrument) were used to create a general main structure of both spectral libraries. The selection of the final pre-treatment, spectral range, threshold values and discrimination method was done based on the number of confused samples generated by each combination of tested algorithms and parameters. The selected combination was in both cases the one that generated the smaller number of confused samples. For samples whose spectral similarities made the classification only with the general structure hard, a further qualification strategy -cascading sub-libraries- was employed.

b) Reference characterization

Seized samples were characterized by means of high-resolution mass spectrometry (HRMS), gas chromatography-mass spectrometry (GCMS) and nuclear magnetic resonance (NMR). All these analyses were carried on at the Department of Analytical Chemistry, University of Valencia, Spain, by Professor Sergio Armenta and collaborators.

HRMS was conducted with electrospray ionization (ESI) on a TripleTOF™ 5600 LC/MS/MS System from AB SCIEX (Redwood City, CA, USA). Mass spectra were recorded using the direct infusion experiment in the positive ion and high sensitivity mode under the following conditions and settings: ion source gas, nitrogen; ion source gas 1 and 2 pressures, 35 and 35 psi, respectively; curtain gas pressure, 25 psi; source gas temperature, 400 °C; ion spray voltage, 5500 V. The AB SCIEX PeakView software was employed for data treatment to obtain accurate mass measurements and isotopic patterns.

Sample solutions were introduced in the system being dissolved in methanol/10 mM ammonium formate in water (80:20 %, v/v) at a flow rate of 0.1 mL/min.

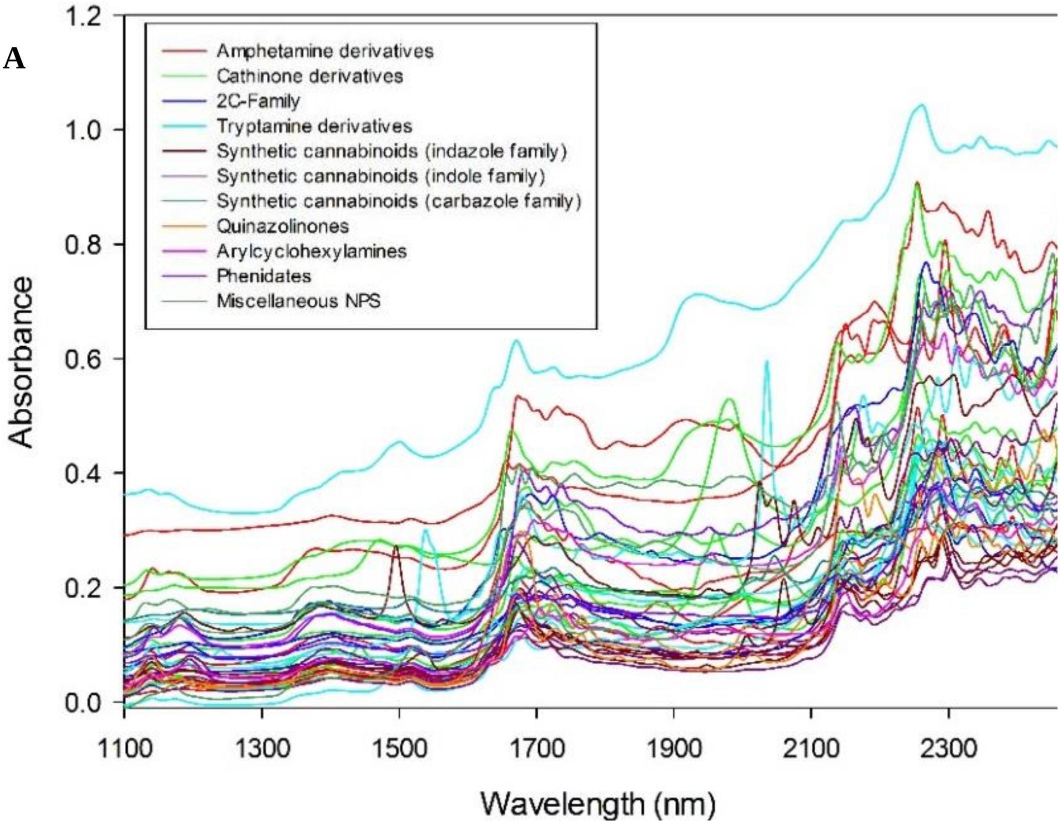
GC-MS was conducted on a 7890A GC system (Agilent Technologies, Santa Clara, CA, USA), equipped with a Zebron ZB-5MS capillary column (30 m × 0.32 mm i.d., film thickness 0.25 µm) and a 5975Cinert XL EI/CI MSD triple axis single quadrupole detector (Agilent Technologies) was used for the identification of the target compound in the sample. Samples were dissolved in acetone and 1 µL solution was injected in the splitless mode at 250 °C, employing helium as carrier gas in constant flow mode at 1 mL/min. Oven temperature program was 150 °C, held for 1 min, increased at a rate of 10 °C/min up to 250 °C, and finally held 5 min. Transfer line and ion source temperatures were 300 and 250 °C, respectively, and an electron voltage at 70 eV was employed for electron ionization. Full scan determinations were performed using the range from 40 to 300 m/z.

NMR spectra of samples dissolved in $CDCl_3$ were acquired at room temperature on a Bruker AVIII spectrometer, equipped with a 5 mm direct probe (Bruker, Billerica, MA, USA). The ^1H spectra were acquired at 300.13 MHz, 298 K with a direct observation, 30° pulse and 16 scan, and TRAF resolution enhancement was applied without line broadening. Chemical shifts (δ ppm) were referenced to tetramethylsilane.

3.3 Results and discussion

The main difference between the data sets obtained from the two different instruments is the spectral range. As can be seen in Figures 1A and 1B, the bench-top instrument enables more NIRS overtones and combination bands than the hand-held instrument, because of the wider spectral range that its detection system makes available. It can be also be observed that the difference in the wavelength intervals employed during the acquisition (2nm in the bench-top and 6.25 nm in the hand-held instrument) has a slight effect on the final resolution of the spectra. Based on these differences, the main challenge in this work was to generate spectral libraries with the same classification target, starting from data sets with the same number of samples but different number of variables.

Furthermore, Figure 1C illustrates the higher absorption intensities in spectra from the hand-held instrument than in spectra from the bench-top instrument, by means of the plot of spectra from the two instrument for the same sample, 2-FMA. These higher values can be understood considering the differences in the radiation sources between the two instruments. The spectral variability between the same families of compounds can be also observed in this set of absorbance spectra, which represents the relevant source of variability provided by differences in substituents of the molecular structures that define each NPS family. However, it is worth mentioning that particularly for cathinone derivatives, the similarities of their NIR spectra forced the construction of particular sub-libraries for classification of samples with slighter molecular differences, as will be exposed below.



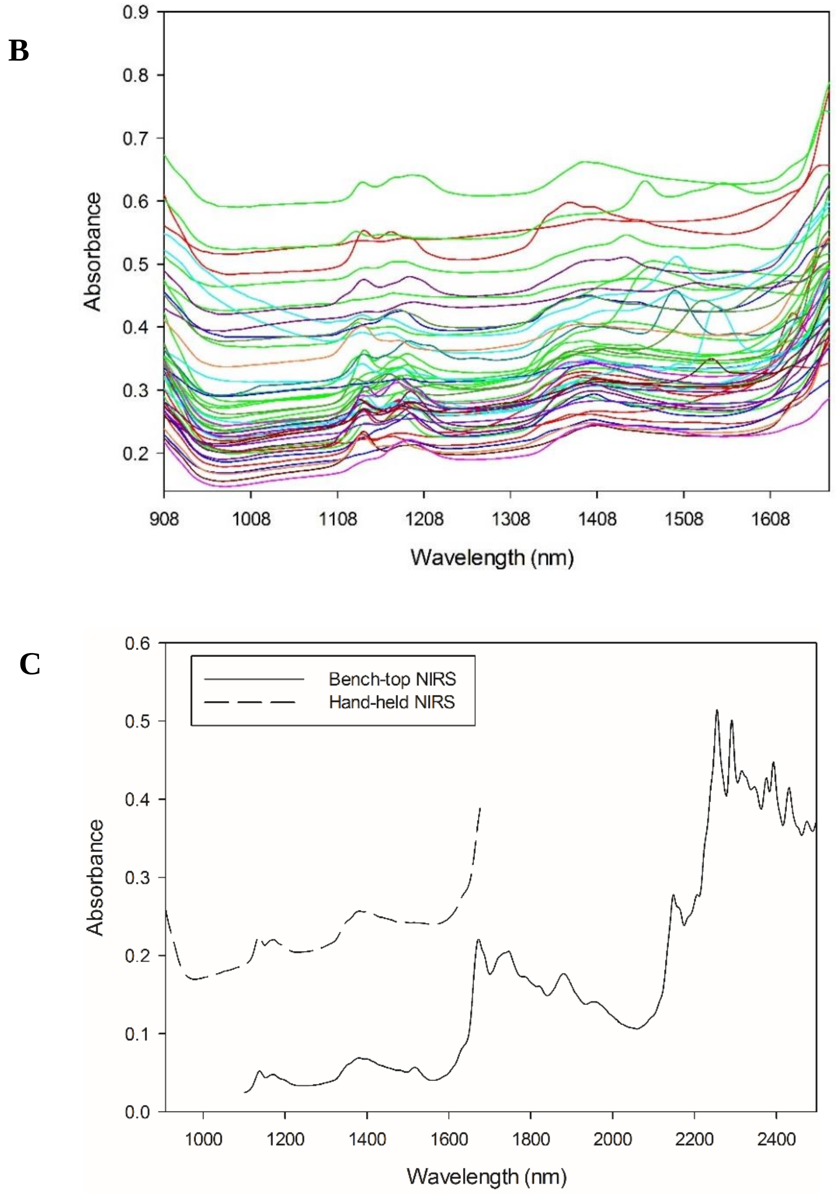


Figure 1. NIR absorbance spectra acquired in reflectance mode for all the families of NPS with ***(A)*** the bench-top and ***(B)*** the hand-held instrument. The graph ***(C)*** illustrates the differences in absorbance intensity and wavelength range for the sample 2-FMA.

3.3.1 Library constructed using data acquired with the bench-top instrument

The first level of the library built with the data obtained from the bench-top instrument, employed the SNV pre-treatment over the spectral range of 1100-2496 nm and Euclidean distance as discrimination method. This general library required the application of the cascading strategy, by means of the creation of 6 sub-libraries in the second level of discrimination, and one sub-library in the third level. The requirement of these additional qualification levels was driven by the similarities found both between samples of the same families of compounds and between families with structural similarities.

Figure 2, illustrates an example of the absorbance spectra acquired using the bench-top NIR for samples of the same family of compounds. This set of samples shows the similarities of combination bands and overtones due to vibrations of common bonds from the main structural framework of amphetamines (aromatic ring and N-H bond of secondary aliphatic amine), as well as the differences provided by their diverse substituents at R_1', R_2, R_3 and R_4. Figure 2, also makes evident the relevant effect of the electronegativity of the halogen substituent in the definition of the aromatic overtones in the range 2100 - 2500 nm.

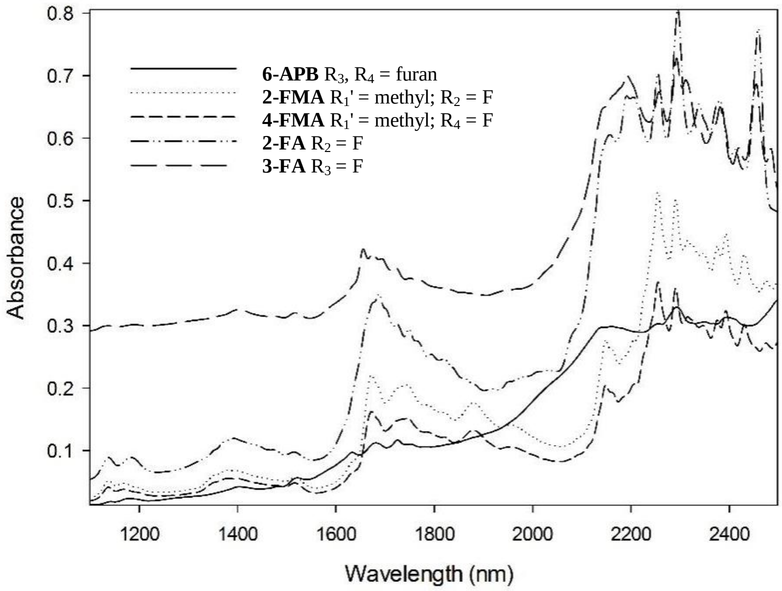


Figure 2. Absorbance NIR spectra of some of the amphetamine derivatives analysed using the bench-top instrument.

Even when SNV pre-treatment emphasized the differences between similar spectra, the Factorization algorithm, based on the calculation of Mahalanobis distances, was required for discrimination of confused samples by means of the sub-libraries described in Table 2. Figure 3 encompasses the scores plots of such sub-libraries, displaying the selected factors in each case.

Table 2. Description of parameters of the sub-libraries created for discrimination of samples ambiguously identified in the first level of the model developed for the data obtained using the bench-top instrument. The number of smoothing points selected for derivative pre-treatments was 9 in all the cases.

Families of samples involved	Sub-cascading level	Spectral range (nm)	Data pre-treatment	Included samples	Threshold
(A) Cathinone derivatives-cannabinoids	2	1626-2402	2D + SNV	4-MePPP	0.01
				JWH-081	0.01
(B) Amphetamine derivatives-miscellaneous	2	1598-2470	2D + SNV	2-FA	0.001
				2-FMA	0.001
				Methiopropamine	0.0007
(C) Cathinones	2	1744-2118	1D + SNV	4-MMC	0.0004
				3-MMC_v1	0.003
				3-MMC_v2	0.0005
(C') Cathinones	3	1896-2124	1D + SNV	3-MMC_v1	0.3
				4-MMC	0.03
(D) Cathinone derivative-T22	2	1120-2484	1D + SNV	Methylone	0.02
				T22	0.002
(E) Miscellaneous-indole cannabinoids	2	1100-2498	1D	3-fluorophenmetrazine	0.02
				MMB-CHMICA	0.007
(F) Tryptamine derivatives	2	1100-2498	1D	α-PVP	0.3
				Butylone	0.02

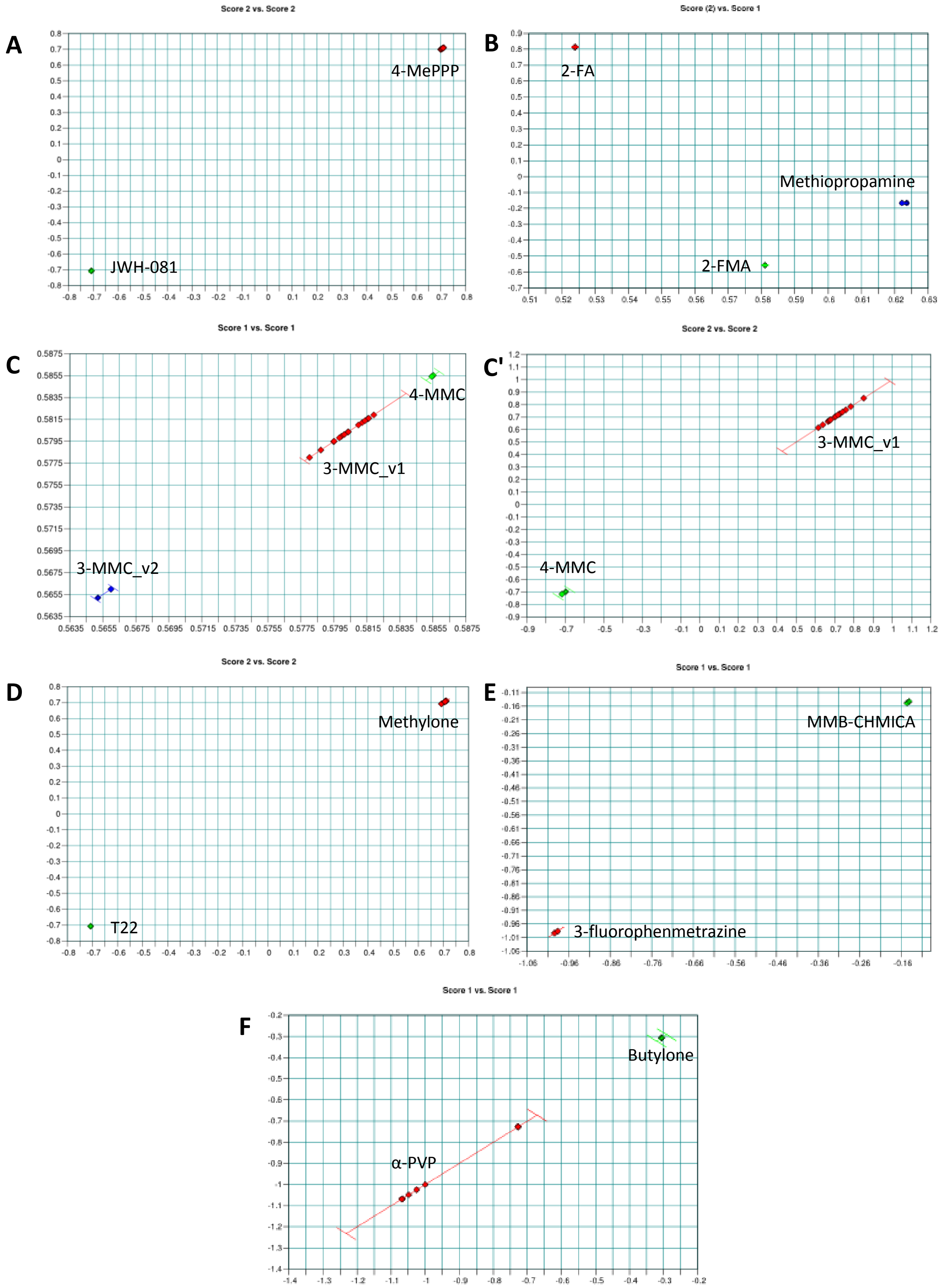


Figure 3. Scores plots of the sub-libraries created in the library constructed using data acquired with the bench-top instrument, for discriminating between ***(A)*** 4-MePPP and JWH-081; ***(B)*** 2-FA, 2-FMA and Methiopropamine; ***(C)*** 4-MMC, 3-MMC_v1 and 3-MMC_v2; ***(C')*** 4-MMC and 3MMC_v1 (created inside of sub-library ***C***); ***(D)*** Methylone and T22; ***(E)*** MMB-CHMICA and 3-fluorophenmetrazine; ***(F)*** α-PVP and Butylone.

As can be seen in Table 2 and Figure 3, in this case, the cathinone derivatives were the compounds that required more efforts for reaching a suitable classification model. Besides the fact of been the family with the higher number of samples in the studied set, relevant differences between the NIR spectral profile of 3-MMC samples motivated the creation of two separated groups: 3-MMC_v1 and 3-MM_v2. The final "_v" has been employed to indicate that different sources of variability were detected in this sample set, and the numbers 1 and 2 were employed for distinction between them. These differences could be explained by different places and dates of seizing. The discrimination between these two sources of variability was possible using the bench-top NIRS instrument.

The spectral library constructed allowed the unambiguous identification of all the samples analysed. The validation of the whole library was completed based on the selectivity (S) values calculated by the Opus software. The S is obtained by means of the ratio of the distances of the average spectra D and the sum of the radii of the closer clusters (thresholds, T_1 and T_2), according the equation:

$$S = \frac{D}{(T_1+T_2)} \qquad \textit{Equation 1}$$

The values of S are useful to understand the capability of the classification model for avoiding confusions between similar samples. The interpretation of S can be done as: $S < 1$: overlapping; $S = 1$: cluster in contact, but without any sample in the area that is in contact; $1 < S < 2$ cluster separated by the minimum possible distance; $S \geq 2$ cluster broadly separated [18]. The selectivity of the library after applying the complete cascading strategy, expressed in percentages of S, regarding all the groups involved in the model was of 1.7% for $1 < S < 2$ and 98.3% for samples with $S \geq 2$.

3.3.2 Library constructed using data acquired with the hand-held instrument

Because of the reduced spectral range available from the portable instrument, regarding the range available from the bench-top instrument, more confusions were found in this case from all the evaluated combinations of discrimination methods, data pre-treatments and thresholds. This spectral set also required a derivative pre-treatment besides the SNV for the description of the general structure, to emphasize the differences acquired with the miniaturized instrument. Additionally, it was required a more branched cascading structure than the one created for the library of the bench-top instrument, as will be showed below. This can be understood considering the lack of bands characteristics of aromatic rings in the missing spectral range (1670-2500 nm). After assessing many diverse combinations, the one able to uniquely identify all the samples studied was the one created using the 2D with 9 points of smoothing window, followed by a SNV pre-treatment, in the spectral range of 908-1676 nm and with Euclidean distance as discrimination method. Afterwards, two sub-libraries internally subdivided were required. Table 3, describes the details of the internal structure of the spectral library built with data acquired using the hand-held instrument.

Table 3. Description of parameters of sub-libraries created for discrimination of samples ambiguously identified in the first level of the library developed using the data obtained with the portable instrument.

Families of samples involved	Sub-cascading level	Spectral range (nm)	Data pre-treatment	Included samples	Threshold
(A) Cathinone derivatives-Miscellaneous	2	1069-1236	1D (9 pts)	3-FMC	0.09
				3-fluorophenmetrazine	0.02
				Butylone	0.2
				Methylone_v1	0.2
				Methylone_v2	0.09
				Methylone_v3	0.02
(A') Cathinone derivatives	3	908-1676	SNV	Methylone_v2	0.03
				Methylone_v3	0.2
(B) Amphetamines-Cathinones-Arylcyclohexyldamines	2	1069-1205	1D (9pts) + SNV	2-FMA, 4-FMA	0.06
				3-MMC_v1	0.4
				4-MEC	0.03
				4-MePPP	0.3
				Etylphenidate	0.03
				MDPV	0.4
				Methoxetamine	0.03
				□-PHP	0.02
				T5	0.008
				PV9	0.05
				3-FEA	0.01
				Threo-4-fluoromethylphenidate	0.09
				T31	0.02
				T33	0.01
				T34	0.007
				3-methoxymethamphetamine	0.02
				4-MeMABP	0.04
				Ethcathinone	0.03
				4-MePPP	0.09
				3-MMC_v2	0.06
				3-MMC_v3	0.1
	3	1075-1261	1D (9pts)	2-FMA, 4-FMA	0.08
				3-MMC_v1	0.3

(B') Amphetamines-Cathinones-Arylcyclohexyldamines				3-MMC_v2	0.06
				3-MMC_v3	0.1
				4-MEC	0.2
				4-MePPP	0.4
				Methoxetamine	0.07
				4-MeMABP	0.1
				Etcatinone	0.1
				T31	0.09
(B'') Cathinone derivatives	4	1100-1212	2D (13pts) + SNV	3-MMC_v1	0.4
				3-MMC_v3	0.6
				4-MEC	0.05
				4-MePPP	0.3

The sub-libraries created in the case of the data acquired using the hand-held instrument were also based on the Factorization method. Figure 4 shows the scores plots obtained with the parameters described in Table 4, with the corresponding factors selected in each case.

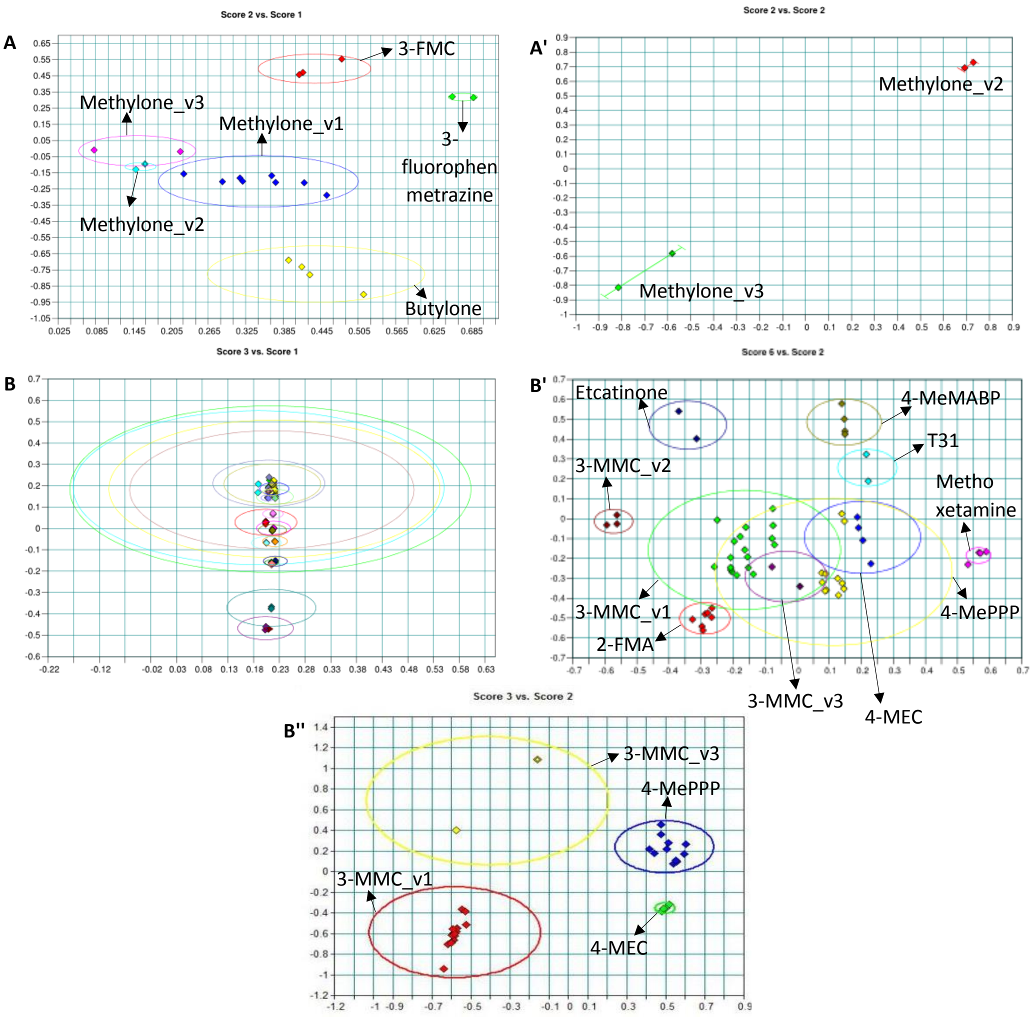


Figure 4. Scores plots of the five sub-libraries created after the first level of discrimination of the model developed using the data from the hand-held instrument. These sub-libraries were required for solving confusions between samples described in Table 4. Letters in this figure corresponds to those written before the names of the families in Table 4.

It can be observed that spectra of 3-MMC showed three sets of samples with differences that were not possible to overcame using the available data pre-treatments. Additionally, the Methylone showed also three sets of differenced samples in the data acquired with the hand-held NIRS instrument. This fact, joined to the similarities of samples from families of Amphetamines, Cathinones and Arylcyclohexyldamines in the spectral range provided for this instrument, compelled the generation of a library more branched than the one developed from the data acquired using the bench-top instrument.

The selectivity of the library after applying the complete cascading strategy, expressed in percentages of *S*, regarding all the groups involved in the model, was of 1.6% of S=1, 23.2% of 1 < *S* < 2 and 75.2% of S ≥ 2.

3.3.3 Comparison of classification models

The main difference between the classification models developed from data acquired using the two instruments are the branches of the structure generated in each case. The library generated from the data acquired with the bench-top NIRS instrument required 6 sub-libraries at the second level of the structure for solving ambiguities between two or three samples in the worst case, and only one of these sub-libraries required an internal sub-library for solving confusions between two samples. The spectral library constructed using the data acquired with the hand-held instrument required 2 sub-libraries in the second level of the structure for solving ambiguities between 6 and 21 samples. In the case of the second of these sub-libraries, two more internal sub-libraries were required for a final unambiguous identification of all the samples involved.

The second relevant difference is related to the percentage of samples classified with *S* values over 2. High distances between clusters is one of the aspects that define the robustness of classification models. For the studied set of NPS samples, the percentage of samples with *S* > 2 is higher for the model developed using data acquired with the bench-top NIRS instrument than for the one created using data from the hand-held instrument. Additionally, the model developed using data acquired with the hand-held instrument displays a 1.6% of samples with clusters in contact, which is not observed in the model of the bench-top instrument. Having clusters in contact is not a desirable situation in classification models, however, complex data sets, as the one created by NPS analysed in a reduced spectral range, can show this kind of situations. The most important in these cases is to demonstrate that none of the samples included into the models is in such overlapping area. Table 4, summarizes all the differences found between the two spectral libraries developed.

Table 4. Comparison of classification models constructed with data acquired with the bench-top and hand-held NIRS instruments.

Assessment criteria	Bench-top instrument	Hand-held instrument
Discrimination method	Euclidean distance	Euclidean distance
Spectral range	1100-2498 nm	908-1676 nm
Data pre-treatment required in level 1	SNV	2D (9pts) + SNV

Number of samples confused in level 1	13	23
Number of sub-libraries in level 2	6	2
Number of sub-libraries in level 3	1	2
Number of sub-libraries in level 4	-	1
Percentage of $S < 1$	-	-
Percentage of $S = 1$	-	1.6
Percentage of $1 < S < 2$	1.7	23.2
Percentage of $S > 2$	98.3	75.2

Even though the spectral library generated using data acquired with the bench-top instrument shows better values of the assessment criteria of the identification models, considering the important reduction of time of detection of illicit drugs that the on-site analysis with the hand-held NIRS instrument can provide, the obtained results confirm that portable instruments could be a valuable tool for early classification of NPS samples during police seizing procedures. Information provided by both instruments can be complementary. The initial identification can be done on-site using the hand-held instrument and such results can be later confirmed and tuned in a short analysis using the bench-top instrument at the laboratory.

3.4 Conclusions

Even when models developed using data from NIRS miniaturized instruments are limited in performance regarding those developed using data provided by bench-top instruments, classification models of NPS based on data from hand-held instruments can be useful to make real-time and on-site decisions that can be confirmed later using high performance analytical instrumentation.

3.5 References

[1] United Nations Office on Drugs and Crime, “Executive summary. Conclusion and policy implications of the World Drug Report 2017,” 2017.

[2] A. Carlsson *et al.*, “Prediction of designer drugs: synthesis and spectroscopic analysis of synthetic cannabinoid analogues of 1H-indol-3-yl(2,2,3,3-tetramethylcyclopropyl)methanone and 1H-indol-3-yl(adamantan-1-yl)methanone,” *Drug Test. Anal.*, vol. 8, no. 10, pp. 1015–1029, 2016.

[3] L. Cornara *et al.*, “Smart drugs: Green shuttle or real drug?,” *Int. J. Legal Med.*, vol. 127, no. 6, pp. 1109–1123, 2013.

[4] S. Strano Rossi *et al.*, “An analytical approach to the forensic identification of different classes of new psychoactive substances (NPSs) in seized materials,” *Rapid Commun. Mass Spectrom.*, vol. 28, no. 17, pp. 1904–1916, 2014.

[5] S. Harkai and M. Pütz, “Comparison of rapid detecting optical techniques for the identification of New Psychoactive Substances in ‘Legal High’ preparations,” *Toxichem Krimtech*, vol. 82, no. Special Issue, p. 229, 2015.

[6] J. Moros, N. Galipienso, R. Vilches, S. Garrigues, and M. De La Guardia, “Nondestructive Direct Determination of Heroin in Seized Illicit Street Drugs by Diffuse Reflectance near-Infrared Spectroscopy,” *Anal. Chem.*, vol. 80, pp. 7257–7265, 2008.

[7] N. Sondermann and K.-A. Kovar, “Screening experiments of ecstasy street samples using near infrared spectroscopy,” *Forensic Sci. Int.*, vol. 106, no. 3, pp. 147–156, Dec. 1999.

[8] K. Tsujikawa *et al.*, “Development of a Library Search-Based Screening System for 3,4-Methylenedioxymethamphetamine in Ecstasy Tablets Using a Portable Near-Infrared Spectrometer,” *J. Forensic Sci.*, vol. 61, no. 5, pp. 1208–1214, Sep. 2016.

[9] I. Baer and R. Gurny, “NIR analysis of cellulose and lactose: Application to ecstasy tablet analysis,” *Forensic Sci. Int.*, vol. 167, no. 2–3, pp. 234–241, Apr. 2007.

[10] R. C. Schneider and K.-A. Kovar, “Analysis of ecstasy tablets: comparison of reflectance and transmittance near infrared spectroscopy,” *Forensic Sci. Int.*, vol. 134, no. 2–3, pp. 187–195, Jul. 2003.

[11] C. Pérez-Alfonso, N. Galipienso, S. Garrigues, and M. de la Guardia, “A green method for the determination of cocaine in illicit samples,” *Forensic Sci. Int.*, vol. 237, pp. 70–77, Apr. 2014.

[12] R. Risoluti, S. Materazzi, A. Gregori, and L. Ripani, “Early detection of emerging street drugs by near infrared spectroscopy and chemometrics,” *Talanta*, vol. 153, pp. 407–413, Jun. 2016.

[13] S. Materazzi, G. Peluso, L. Ripani, and R. Risoluti, “High-throughput prediction of AKB48 in emerging illicit products by NIR spectroscopy and chemometrics,” *Microchem. J.*, vol. 134, pp. 277–283, Sep. 2017.

[14] K. Tsujikawa *et al.*, “Application of a portable near infrared spectrometer for presumptive identification of psychoactive drugs,” *Forensic Sci. Int.*, vol. 242, pp. 162–171, Sep. 2014.

[15] C. G. Pederson *et al.*, “Pocket-Size Near-Infrared Spectrometer for Narcotic Materials Identification,” in *Proceedings of SPIE—The International Society for Optical Engineering*, 2014, p. 9101.

[16] M. Blanco and M. A. Romero, “Near-infrared libraries in the pharmaceutical industry: a solution for identity confirmation,” *Analyst*, vol. 126, pp. 2212–2217, 2001.

[17] A. Y. Miró Vera and M. Alcalà Bernàrdez, “Near-Infrared Spectroscopy in Identification of Pharmaceutical Raw Materials,” *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, pp. 1–19, 2017.

[18] Bruker, “OPUS IDENT User Manual, Spectroscopy Software, Version 7.” Ettlingen, 2011.

4. Inline monitoring of recombinant production of Lipase B from Candida antarctica in Pichia pastoris using glycerol as carbon source

4.1 Introduction

The basidiomyceteous yeast *Candida anctartica (C. anctartica)*, produces two different lipases, named A and B. Both lipases are catalysts highly stable in an immobilized form, but it has been demonstrated that Lipase B tolerates important variations in experimental conditions, maintaining particularly efficient biocatalyzing properties. The most relevant of these properties is probably the high degree of substrate regio and enantioselectivity for a great number of different organic reactions, many of them nowadays commercially scaled up. Because of that, Lipase B has been extensively used in the resolution of racemic alcohols, amines and acids, as well as in the preparation of optically active compounds from meso reactants. Optically pure compounds can be certainly difficult to obtain by alternative routes and some of them have an important synthetic value. Lipase B has been also intensively used as a regioselective catalyst to selectively acylate different carbohydrates [1].

Recently, the conditions for the obtention of Lipase B by its recombinant production in *Pichia pastoris (P. pastoris)* -yeast newly named *aka komagataella phaffii*- has been enhanced [2]. *P. pastoris* is a high cell density expression system, probably one of the most promising protein expression systems [3]. High cell density yeast expression systems have been increasingly applied to manufacturing human and mammalian proteins, among other reasons, because of their extremely rapid process dynamic [4]. Following this trend, *P. pastoris* system combines a high growing speed on simple media and important eukaryotic features such as glycosylation. Therefore, *P. pastoris* provides high expression levels in an economical and ease of manipulate system, able to perform complex post-translational modifications. Two types of production are possible for this yeast: inductive and constitutive. Although the last one provides less product, it has the advantage of circumvent the need of methanol use which is a hazard when aiming the scale up of the process[5].

The over or under feeding of *P. pastoris* systems can have serious consequences on the formation/stability of its products. This fact turns the multianalyte real-time monitoring up to be a persistent requirement for this kind of system, even when real-time monitoring tools have been desirable to the bioprocess monitoring world from very earlier efforts [6]. Advances in this task has been meet by means of the development of analytical methods based on *in-situ* sensors [4]. Such in-*situ* sensors have to fulfil very particular requirements, the most relevant are:

- Long-term corrosion stability and biological inactivity.
- Capability of maintaining the asepsis of the process.
- Reaching a wide detection range of the target parameters for covering all their changes during the process.
- Production of a fast response to enable opportune corrective action procedures
- Preserving the integrity and calibration after sterilization conditions (high temperatures and pressures different from ambient air pressure) [7], [8].

In view of these characteristics, optical and spectroscopic techniques can be the base of non-invasive in-situ sensors with a huge potential for this kind of processes. This is basically due to its non-analyte consumption, non-sampling step and no further reagents necessity. For most of the optical and spectroscopic probes, sterilization conditions are easy to overcome because no other internal installations than optical windows are required in their inner arrangements. Additionally, spectroscopic methods do not show any time delay, therefore they can provide information at real-time [9]. Among other techniques as Raman spectroscopy and Optical Density (*OD*), one of the techniques that have been applied to bioprocesses monitoring, including *P. pastoris* cultivations, is Near Infrared Spectroscopy (NIRS) [10]. Applications of NIRS to bioprocesses monitoring can be found on diverse microbial processes, insect cell and animal cell cultures [4]. The evolution of such applications started with more simple systems, tracked under anaerobic conditions and low agitation, to derive to more complex systems with vigorous aeration and agitation [9]. The common pathway for developing analytical methods for bioprocesses monitoring based on NIRS starts with at-line or on-line (bypass or ex-situ) measurements which are later adapted to more challenging in-line (in-situ) conditions [11]. Most of the research reports in this field display studies based on only one mode of NIRS data acquisition, selected considering the particular physical characteristics of the system. In such cases, specific analyte models have been built both for the entire process data [8] and for segments of the data in accordance with the process evolution time (which is related to changes in the spectral response) [12]. Some common aspects of these previous studies are: low to moderate Biomass levels (Biomass concentration range 0-16 g/L), relatively simple matrices (soluble media, frequently chemically defined) and no complex changes in the physical characteristics of the process fluid over the process evolution. In general, from the spectroscopic point of view, the application of NIRS to very high cell density systems (Biomass concentration range 40-100 g/L) is much more challenging, due to the significant effect on the spectral data of the diffuse reflection of radiation generated by solid particles of Biomass. Nevertheless, some previous studies have demonstrated the feasibility of using NIRS for on-line monitoring of processes that employed *P. pastoris* as expression system, which faced the complex changes in the physical characteristics produced in the medium [4], [13], [14].

A valuable advantage of spectroscopic techniques (absorbance or transmittance measurements) compared to simple OD measurements, is the possibility of obtaining more information of the process, in addition to Biomass concentration. Even when Biomass concentration is one of the most critical measurements in bioprocesses [7], substrate and product(s) concentrations, among other parameters, are also desirable results of an analytical method for in-line monitoring of a biotechnological process. NIRS provides key overtones and combination bands in these kind of processes that can be correlated to O-H bonds of alcohols, C-H bonds of aliphatic and aromatic carbon compounds, as well as N-H bonds of proteins [15].

Based on the previously mentioned contributions, and considering that, to the best knowledge of the authors, none totally in-line monitoring method based on NIRS for *P. pastoris* expression systems has been reported so far, the objective of this work is to develop an analytical method based on NIRS for the in-line monitoring of the concentrations of Biomass, Total protein, Glycerol and Nitrogen as well as the Lipolitic activity, during the recombinant production of Lipase B from *C. anctartica* in *P. pastoris* using Glycerol as carbon source. For this purpose, the first step was to complete a feasibility study focused on two aspects: assessing the capabilities of the technique for the specific system to be studied and selecting the proper acquisition mode of the instrument. Afterwards, the development of the quantitative models was sequentially addressed by means of three different sets of samples that encompassed the chemical and physical variability of the bioprocess from a minimum to a high level of inclusion.

4.2 Materials and methods

4.2.1 Bioprocess inoculation and evolution

a) Microorganism

The microorganism used in this study was the wild-type yeast strain *P. pastoris* X-33, previously modified by inserting the vector pPGKΔ3_PRO_LIPB, for constitutive expression of recombinant lipase B of *C. antarctica* (rLipB). The lipase gene was synthesized by the EpochBiolabs in pBKSII vector, using codon optimization for *P. pastoris*. The gene was subcloned in the plasmid pPGKΔ3_PRO using the enzymes XhoI and NotI generating the vector pPGKΔ3_PRO_LIPB that presents the rLipB under control of the PGKΔ3 promoter and a signal peptide α-factor with optimized codons. Besides, the vector has Sh ble gene as selective mark toward zeocin [2]. The carbon source for this microorganism was glycerol.

b) Batch medium

Batch medium used contain per liter: 2.0 g citric acid, 12.4 g $(NH_4)_2HPO_4$, 0.022 g $CaCl_2$.2 H_2O, 0.9 g KCl, 0.5 g $MgSO_4$.7 H_2O and 4.6 ml PTM_1 trace salts solution (6.0 g $CuSO_4$.5 H_2O, 0.08 g NaI, 3.0 g $MnSO_4$. H_2O, 0.2 g Na_2MoO_4.2 H_2O, 0.02 g H_3BO_3, 0.5 g $CoCl_2$, 20.0 g $ZnCl_2$, 65.0 g $FeSO_4$.7 H_2O, 0.2 g biotin and 5.0 ml H_2SO_4). The initial carbon source varies according to 0h of each cultivation, pH was maintained at 7.0 with NH_4OH 15%, temperature at 30°C and atmospheric pressure.

c) Fed-batch medium (starting around 20h)

Fed medium used contain per liter: 550 g glycerol, 10 g KCl, 6.45 g $MgSO_4$.7 H_2O, 0.35 g $CaCl_2$.2 H_2O and 12 ml PTM_1 trace salts solution.

d) Process variables

An applikon Biobundle bioreactor, with 2 L working volume was used. Dissolved oxygen (DO) was maintained at 30% by a cascade control changing stirring between 500-1000 rpm and aeration between 0-1vvm, a mix of compressed air and pure oxygen was provided manually, when needed. Temperature was maintained at 30°C and the working pH was controlled at 7 with 15% v/v NH_4OH. Anti-foam was added during the process as necessary.

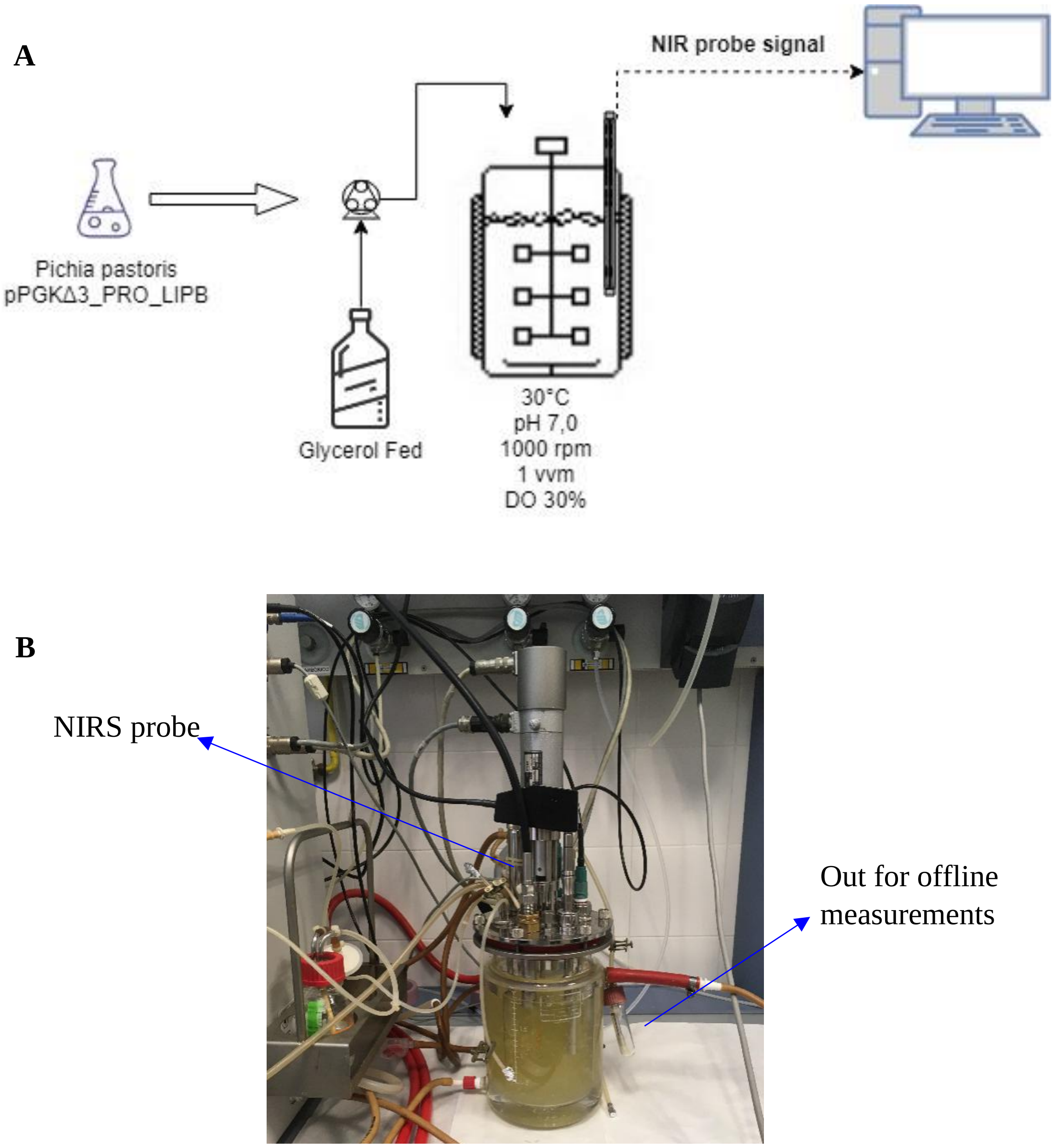


Figure 1. (*A*) General representation of the reactor and conditions of the cultivation process with the in-line connection of the NIRS fiber optic probe; (***B***) picture of the reactor during a process run.

4.2.2 Analytical monitoring

a) Reference Methods per analyte

Biomass concentration: Optical density at 600 nm in a Hach lange GmbH – DR3900 spectrophotometer. The relationship with dry cell weight (DCW) was: DCW (g/L)= 0.3068*abs [16].

Glycerol concentration: HPLC equipment from HP 1050 liquid chromatograph (Dionex Corporation, Sunnyvale, CA, USA) using an ICSep ICE COREGEL 87H3 column (Transgenomic Inc., Omaha, NE, USA) The temperature was maintained at 40°C using as the mobile phase sulfuric acid solution 0.0032 M at flow rate of 0.5 ml/min. The sample volume injected was 20 µL. The Glycerol concentration was determined as the mean of triplicates, Relative Standard Deviation (RSD) <1%.

Total protein concentration: Bradford method [17]. Total protein was determined as the mean of triplicates, RSD < 2%.

Nitrogen concentration: Methodology described by [18], [19]. Ammonium sulfate was used as the standard. Nitrogen concentration was determined as the mean of triplicates, RSD < 2%. Nitrogen data provided refers to ammoniac salts present in the medium and added by pH control.

Lipolitic activity: Titrimetric method, by means of a hydrolysis reaction with 56 mM tributyrin at 40°C and pH 7, using a pHstat [20]. The titrating reagent employed was a 0.06 mM solution of NaOH. One unit of lipolytic activity (U) is defined as the quantity of enzyme needed to catalyse the production of 1 µmol butyric acid (volumetric analysis) per minute under the assay conditions. The activity was determined as the mean of triplicates, RSD < 10%.

Additional analysis related to the stability of rLipB at different pH and temperature ranges, as well as the substrate specificity and further characterization of rLipB were done following the methodologies described by [2].

All the reference values were determined by personnel of the School of Chemical Engineering of the *Universitat Autònoma de Barcelona*, properly trained for such purpose.

b) NIRS instruments

Feasibility study: The instrument employed for preliminary evaluations was a NIR spectrometer from Foss (Denmark), model 5000, fitted with a Rapid Content Analyzer model 6500, used in reflectance and transflectance acquisition modes. Each spectrum was the result of 32 scans acquired with a spectral resolution of 10 ± 1 nm and a data interval of 2 nm, in the wavelength range from 1100 to 2498 nm. For acquisitions in reflectance mode, a quartz cell of 4 cm of diameter was employed, and for transflectance mode a reflector of gold with 1 mm of path length was added. All the spectral data was acquired at room temperature. The software Vision, version 2.51 was employed for the data acquisition.

Models development: For offline and inline measurements, the equipment employed was a FT-NIR spectrometer from Bruker (Billerica, MA, USA) model Matrix F, with fiber optic coupling for immersion probe and a TE-InGaAs detector. The fiber optic probe was steam-sterilizable and its acquisition mode Transflectance, with a path length of 10 mm. Each spectrum was acquired as an average of 512 scans with wavenumber intervals of 16 cm^{-1}, in the spectral range from 12000 to 4300cm^{-1}. Off-line measurements were conducted at room temperature. In-line measurements were acquired at 30°C and every 3 minutes over the entire process experiences recorded using this data collection mode. The software Opus, version 7.5 was employed for the data acquisition.

NIRS data analysis (development of the models or modelling) was made based on principal component analysis (PCA) and partial least squares (PLS) regressions, accomplished using Unscrambler from CAMO, version 10.3, and Solo, from Eigenvector Incorporated, version 8.2.1.

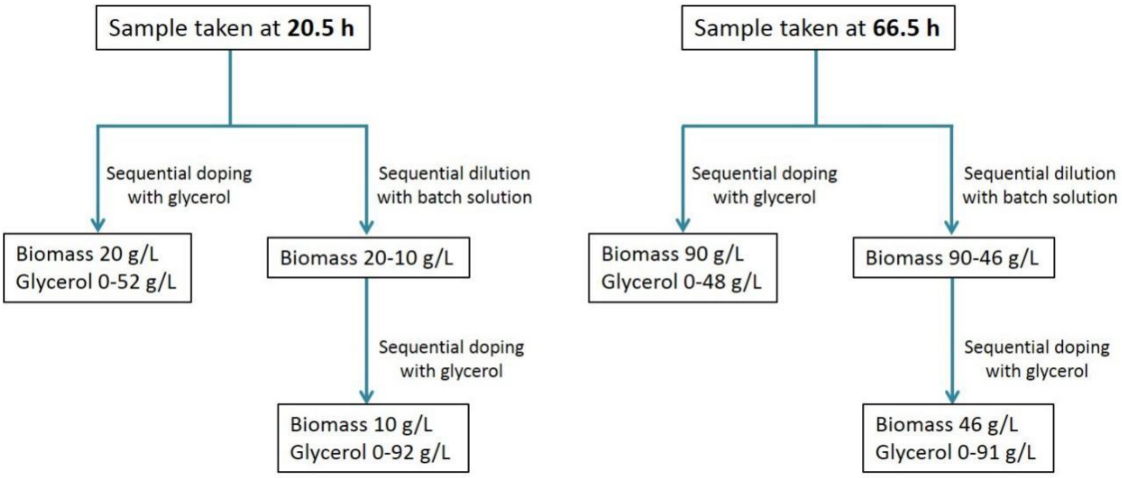
4.3 Feasibility study

Previous contributions state that the presence of water is not an impediment for using NIRS as a monitoring tool for bioprocesses [21]. Due to the changes that intermolecular hydrogen bonds of water undergo by the presence of O-H and N-H bonds from alcohols and protein molecules, combination and overtone bands of the NIR spectrum of water are modified. However, considering that due to the important dipolar moment of the O-H bond, water is a strong absorber in the Infrared region -including the NIR- a feasibility study was set for assessing the advantages and disadvantages of applying NIRS to the particular conditions of the process intended to be studied. Additionally, since the progressive increment of solid particles of Biomass in the liquid medium over the process evolution, the selection of the appropriate acquisition mode required, in this case, a previous evaluation, due to such increment has remarkable light scattering effects on the NIRS data. This fact prompted the second objective of this preliminary study: the selection of the appropriate acquisition mode. The development of this task considered previous reports on NIRS methodologies for at-line, on-line and in-line monitoring of bioprocesses, and those results pointed out that data collection is a relevant issue for generating the best possible models for this kind of applications. Regarding this aspect, it is important to indicate that Arnold et al. overcame the changes in viscosity observed during the advance of a process comprising a filamentous microorganism (*Streptomyces* *fradiae*), by segmenting the spectral data based on the physical characteristics of the matrix, which allowed more accurate PLS models for prediction of tylosin concentrations using at-line NIRS measurements [22]. On the other hand, Crowley et al. proposed the use of two acquisition modes for the also at-line monitoring of a recombinant *Pichia pastoris* fed-batch bioprocess. They used the transmission acquisition mode for the first part of the process (when the concentration of Biomass comprised the range 0-64 g/L) and the reflectance acquisition mode for monitoring the second part of the process (Biomass concentration 64-80 g/L) [4]. Alternatively, Finn et al. avoided the dominance of the whole matrix spectrum by Biomass increment using filtrate samples and semi-synthetic filtrate samples. This strategy favoured the development of at-line monitoring methods for successful quantification of ethanol and glucose in a *Saccharomyces cerevisiae* fed-batch bioprocess [23]. Additionally, Tamburini et al. developed on-line and in-line NIRS methods for monitoring homolactic (using *Lactobacillus casei*) and heterolactic (using *Staphylococcus xylosus*, *Lactobacillus fermentum* and *Streptococcus thermophilus*) cultivations. In this work they state the successful application of reflectance acquisition mode for the homolactic process (Biomass concentration 0-16 g/L) and the transflectance mode for the heterolactic process (Biomass concentration 0-16 g/L, but heavy aeration and agitation conditions). The choice of an aerobic bioprocess was made particularly for investigating the effect on the spectral signal of heavy aeration and agitation conditions, which showed that both factors have a clear impact on the NIRS data [8]. It is worth mentioning that all these contributions are based on the use of NIRS instruments with lower analytical features (filters or LED systems) than the instrument intended to be used for the present work, a Fourier transform (FT)-NIR spectrometer.

FT-NIR spectrometers are based on the use of a permanently aligned interferometer, capable of recovering the intensities of individual wavelengths in the NIRS region with the best characteristics currently available in terms of wavelength precision and accuracy, signal to noise ratio and scan speed [24]. Based on these characteristics, the acquisition mode was studied to ensure taking as much advantage as possible of this instrument capabilities. This feasibility study was based on the analysis of samples prepared at the analytical laboratory, with the objective of creating a set of spectral data able to show the relationships between changes in Biomass and Glycerol concentrations and absorbance and wavelength values. The strategies followed for preparing those samples were doping with Glycerol and dilution with batch medium of two samples from a bioprocess monitored both with the reference methods and NIRS off-line measurements. Due to the cultivation studied involved two different stages: a batch stage and a fed-batch stage, the first of the samples employed for this preliminary study was taken at the end of the batch stage (20.5 h), and the second one at the end of the fed-batch stage (66.5 h). Both identification times are counted from the inoculation starting point, which is taken as 0 h.

With the aim of generating samples with known concentrations of Glycerol at diverse Biomass concentration levels, both samples were divided into two portions. The first portion was doped by sequential addition of Glycerol in the range 0-52 g/L for the sample taken at 20.5 h and 0-48 g/L for the sample taken at 66.5 h. This concentration range was established based on the real concentration values of Glycerol that could be needed to monitor during the cultivation processes. Changes in Glycerol concentration are of particular interest due to the fact that the growing of *P. pastoris* is clearly controlled by the presence or absence of such carbon source [2], as was mentioned before.

The second portion of the samples was diluted first by sequential addition of batch solution in the Biomass concentration range of 20-10 g/L for the sample taken at 20.5 h and 90-46 g/L for the sample taken at 66.5 h. The NIRS data acquisition was done as soon as the doping or diluting substance was added to the corresponding portion of the samples. Finally, the portions of the samples diluted with batch solution were also doped with Glycerol, but in this case in a wider concentration range: 0-92 g/L of Glycerol for the sample diluted at 10 g/L of Biomass and 0-91 g/L of Glycerol for the sample diluted at 46 g/L of Biomass. The entire process of samples preparation and spectra acquisition was completed at room temperature. Scheme 1, summarizes the sample preparation procedure for this feasibility study:


Sample taken at **20.5 h**
Sequential doping with glycerol
Sequential dilution with batch solution
Biomass 20 g/L
Glycerol 0-52 g/L
Biomass 20-10 g/L
Sequential doping with glycerol
Biomass 10 g/L
Glycerol 0-92 g/L
Sample taken at **66.5 h**
Sequential doping with glycerol
Sequential dilution with batch solution
Biomass 90 g/L
Glycerol 0-48 g/L
Biomass 90-46 g/L
Sequential doping with glycerol
Biomass 46 g/L
Glycerol 0-91 g/L

Scheme 1. Flow followed for the preparation of samples for the feasibility study.

Figure 2, shows the effect on the NIRS absorbance data at room temperature of diluting the samples with batch medium –i.e. Biomass concentration decreasing- in reflectance mode (Figure 2A) and transflectance mode (Figure 2B). As can be observed, the presence of Biomass particles can be clearly detected by the technique. The effect in absorbance of those solid particles is more remarkable in the baseline shift of the spectra acquired in reflectance mode than in those acquire in transflectance mode. However, considering that the drift of the instrument is the same in both measurements, using reflection as acquisition mode provides spectra of samples with Biomass concentration under 20 g/L with a high level of noise in the range 1400 – 2500 nm. The spectra resulting from the analysis using transflectance mode show less noise over the whole spectral range and provides a profile with less baseline effect driven by Biomass concentration, and more representative of the water spectrum profile. This is probably due to the reflector of gold employed to return the diffuse reflectance of the particles of Biomass to the detector. Additionally, spectra acquired using transflectance illustrate a distribution of the intensity of absorbance units -in the ranges 1496-1892 nm and 2130-2368 nm- directly proportional to the Biomass increment, as can be seen in Figure 2B.

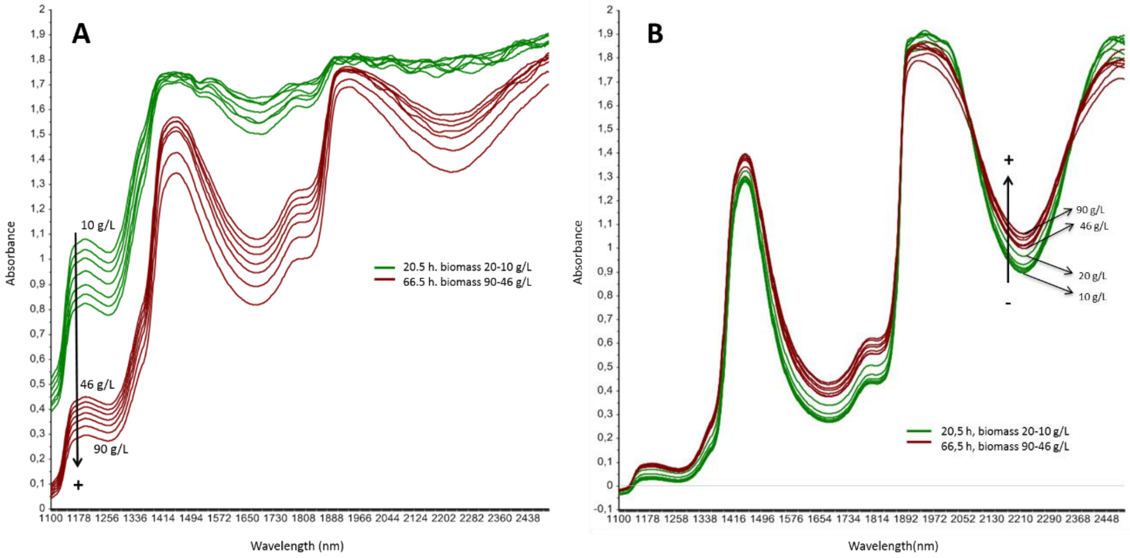


Figure 2. NIR spectra of samples diluted with batch solution. Sample taken at 20.5 h diluted in the range 20-10 g/L and sample taken at 66.5 h diluted from 90-46 g/L, acquired in (***A)*** reflectance and (***B***) transflectance modes.

The effect of doping with Glycerol on samples with different Biomass concentrations at room temperature and without changes in other process variables was studied by means of results shown in Figure 3.

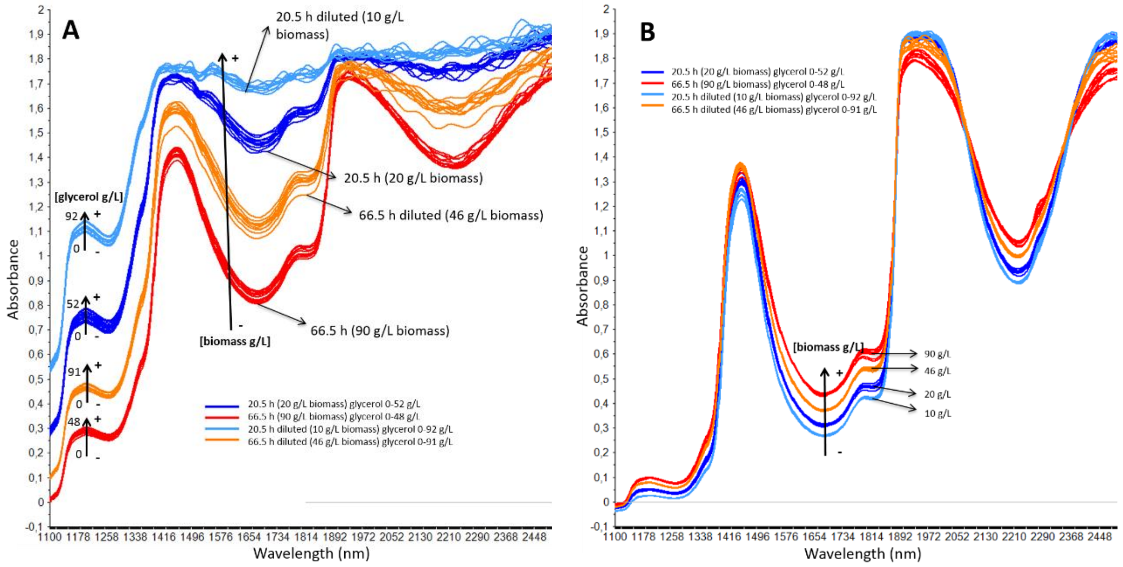


Figure 3. NIR spectra of sample of 10 g/L of Biomass doped with Glycerol in the range 0-92 g/L, sample of 20 g/L of Biomass doped with Glycerol in the range 0-52 g/L, sample of 46 g/L of Biomass doped with Glycerol in the range 0-91 g/L and sample of 90 g/L of Biomass doped with Glycerol in the range 0-92 g/L; acquired in (***A)*** reflectance and (***B)*** transflectance modes.

The effect on the NIRS data of changes in the Biomass concentration previously observed was confirmed with this new set of samples. Nevertheless, the spectral data acquired using the transflectance mode showed not only a better signal to noise ratio over the whole recorded spectral range, but also provided a particular band in the range 2232-2340 cm^{-1}, which was not detected before. Figure 4, shows a zoom of this area, where can be observed that the increment of the absorbance intensity is directly proportional to the concentration of Glycerol, at least at these conditions of minimum variability in the system.

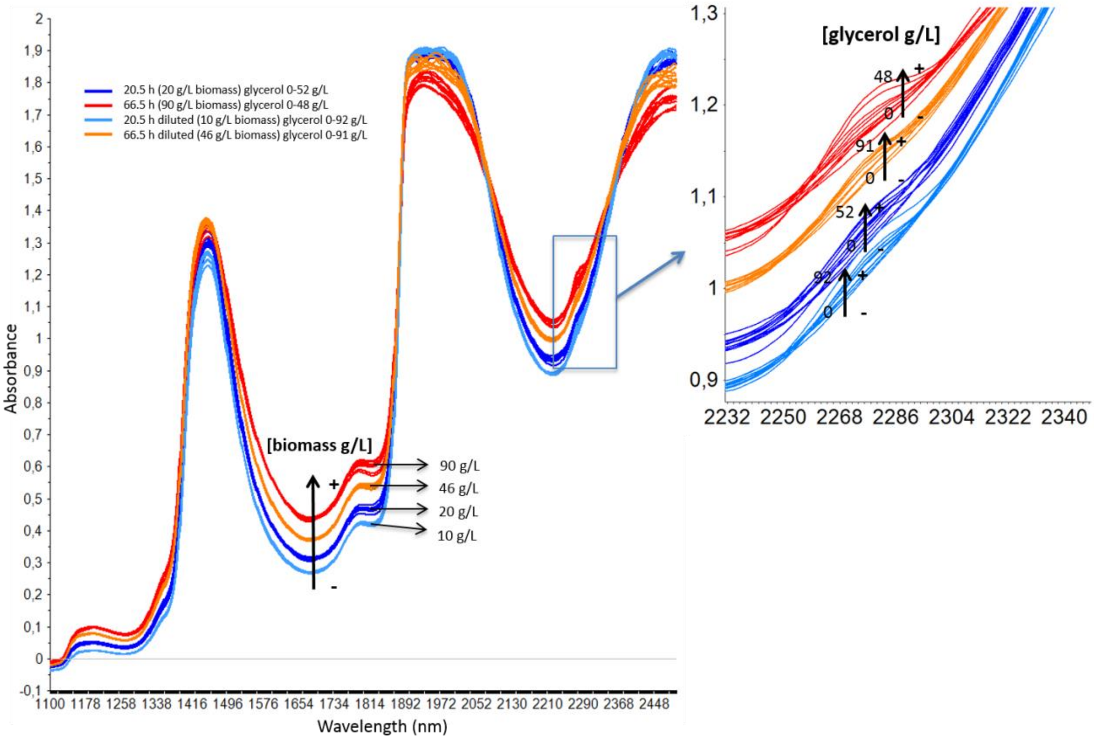


Figure 4. Spectra of sample of 10 g/L of Biomass doped with Glycerol in the range 0-92 g/L, sample of 20 g/L of Biomass doped with Glycerol in the range 0-52 g/L, sample of 46 g/L of Biomass doped with Glycerol in the range 0-91 g/L and sample of 90 g/L of Biomass doped with Glycerol in the range 0-92 g/L, acquired in transflectance mode, with zoom in the range 2232-2340 nm.

A PCA of the absorbance data acquired by means of the two acquisition modes, displayed in a clearer way the differences described before. Even when the first principal component is dominated by the Biomass concentration in both cases (Figure 5), in the scores plot of the data acquired in transflectance mode it is possible to observe the distribution of the samples along the second component according to the Glycerol concentration (Figure 5B).

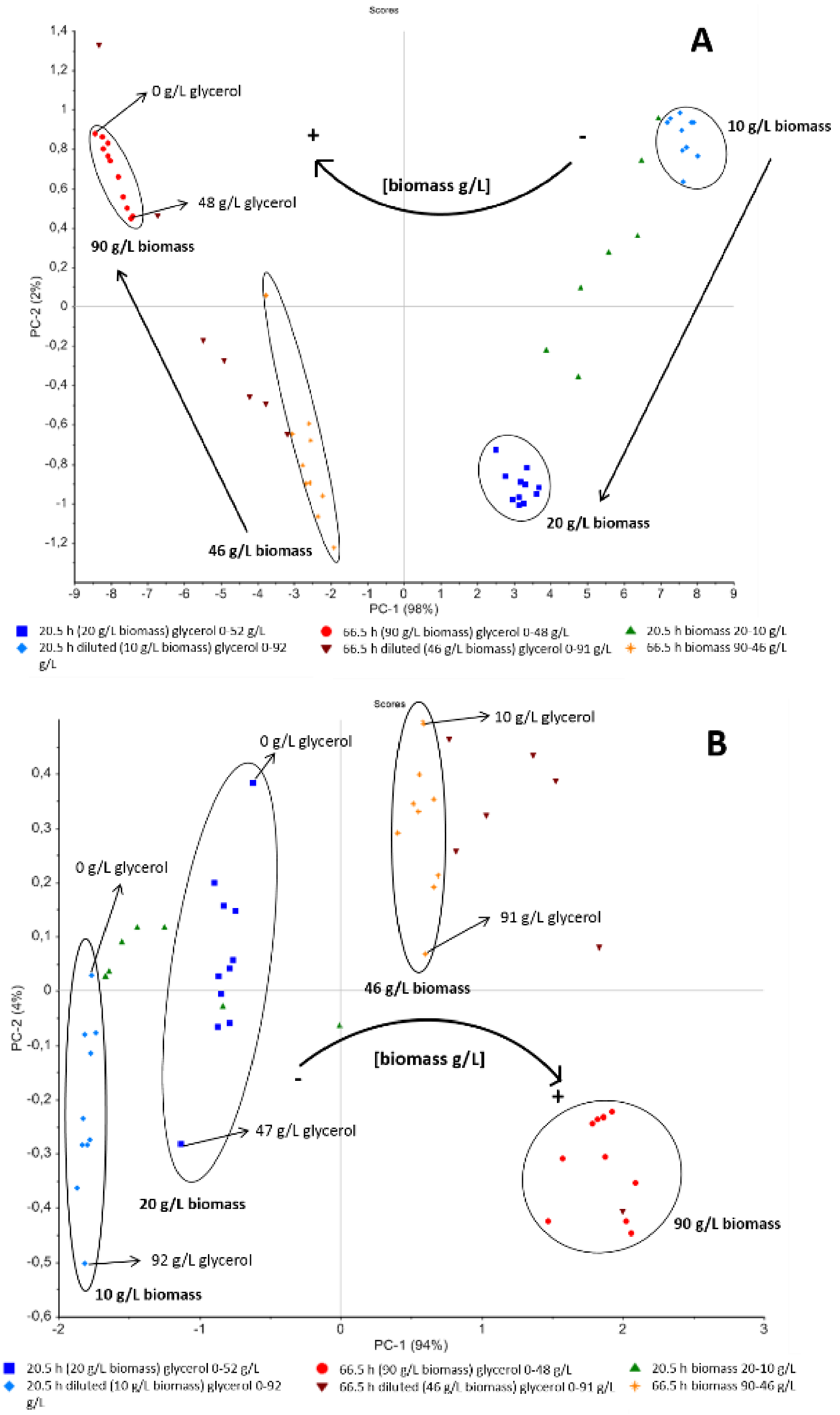


Figure 5. Scores plots of PCA of NIRS data in absorbance of samples recorded in (***A***) reflectance and (***B***) transflectance acquisition modes.

Based on these results, and as a final tool for supporting the selection of the acquisition mode, PLS regressions were calculated for Biomass and Glycerol using the data from the two acquisition modes. For Biomass, it was possible to build a model from the reflectance data using the whole spectral range, i.e. 1100-2500 nm, and without any mathematical spectral data pre-treatment. This model required 2 latent variables (LV) or PLS factors for explaining a high percentage of the variability (over 97 %), generated a root mean square error of calibration (RMSEC) of 5.16 g/L and a coefficient of determination of the regression, R^2 = 0.97. It was also possible to build a model based on the transflectance data using the same spectral range and none data pre-treatment, that required 2 factors also, but generated a RMSEC of 4.36 g/L and a R^2 of 0.98.

Additionally, for Glycerol the same calculations were done. In this case, it was necessary to create two different models for the data acquired in reflectance, because it was not possible to obtain a linear correlation between the spectral data and the reference values available for this analyte in the entire concentration range of interest for Biomass. The data pre-treatment employed in this case was the second derivative Savitzky-Golay (2D) followed by the standard normal variate (SNV), in the spectral range of 1100-1290 nm, the very first part of the spectra, where the signal to noise ratio is still in a proportion useful to obtain valuable information. The model created for the concentration of Biomass of 20 g/L had a RMSEC of 1.89 g/L and a R^2 of 0.98. The model created for the concentration of Biomass of 90 g/L had a RMSEC of 1.72 g/L and a R^2 of 0.99. In both cases the first factor explained the maximum percentage of the variability of the studied data. On the other hand, using the data acquired using the transflectance mode, it was possible to calculate one unique regression for Glycerol over the whole concentration range of interest for Biomass. The data pre-treatment in this case was 2D (with 11 points of smoothing window, the value employed over the whole work for derivative pre-treatments), using the spectral range of 2220-2320 nm, and also only one factor was required for the calculation. Table 1 recapitulates the results of PLS calculations of this feasibility study.

Table 1. Figures of merit of the preliminary PLS regressions constructed with the data from the feasibility study.

Analyte	Reflectance					Transflectance				
	LV	PT	SR (nm)	RMSEC (g/L)	R^2	LV	PT	SR (nm)	RMSEC (g/L)	R^2
Biomass (g/L)	2	None	1100-2500	5.16	0.97	2	None	1100-2500	4.36	0.98
Glycerol (g/L) at 20 g/L of Biomass	1	2D+SNV	1100-1290	1.89	0.98	1	2D	2220-2320	2.89	0.97
Glycerol (g/L) at 90 g/L of Biomass	1	2D+SNV	1100-1290	1.72	0.99					

LV: Latent Variables (PLS factors); PT: Data Pre-treatment; 2D: Second Derivative Savitzky-Golay; SNV: Standard Normal Variate; SR: Spectral Range; RMSEC: Root Mean Square Error of Calibration expressed; R^2: Coefficient of determination for the PLS regression.

Based on these results, it is possible to state that: (1) off-line NIR spectral changes due to changes in Biomass in the concentration range of 10-90 g/L and Glycerol in the concentration range 0-92 g/L at room temperature; (2) it can also be used for calculating preliminary PLS regressions models for both analytes. Under the studied conditions, spectra in reflectance and transflectance acquisition modes are useful for determination of Biomass in the mentioned ranges. However, for the determination of Glycerol, the transflectance mode enables a simpler modelling strategy.

4.4 Results and discussion

As a consequence of the feasibility study, transflectance was employed as acquisition mode. The development of the models was done by the progressive study of the sources of variability of the process and their successive inclusion in PLS regressions. For this aim, a succession of three steps was traced in this work. The first step, involved modelling using only samples prepared at the analytical laboratory (not from a cultivation experiment), which included only part of the chemical variability of the process and minimized as much as possible the real physical and chemical variability. At this point, preliminary PLS models were calculated for Biomass and Glycerol. The second step was the partial inclusion of the process variability into such models by means of NIRS data obtained from bioprocess monitored off-line. The final step encompassed the development of models for prediction of the concentration of four analytes -Biomass, Glycerol, Total protein and Nitrogen- and one process parameter -Lipolitic activity- based on NIRS data acquired in-line. Results of these three steps are discussed in the next headlines.

4.4.1 Samples prepared volumetrically

The most basic chemical variability of the process was studied by means of the volumetric preparation of samples at key concentrations of the main analytes: Glycerol, Biomass and Total protein. Syringes of ± 0.1 mL of precision were employed for creating two sets of samples from stock solutions of Biomass at 200 g/L, Glycerol at 100 % v/v and Total protein at 1000 mg/L. These sets of samples will be labelled as laboratory samples in the subsequent. Both sets of laboratory samples had five fixed concentrations of Biomass in common: 0, 25, 50, 75 and 100 g/L. One set of the samples comprised nine different concentrations of Glycerol at each Biomass concentration level (0, 5, 10, 15, 20, 25, 30, 35 and 40 g/L of Glycerol). The other one, comprised four concentrations of Total protein per Biomass concentration level (75, 150, 225 and 300 mg/L of Total protein). Both sets of samples were useful for assessing the capabilities of the FT-NIRS for detecting changes of main interest chemicals during *P. pastoris* cultivation processes using Glycerol as carbon source, in conditions of minimized physical and chemical variability. Additionally, these samples allowed the evaluation of the most appropriate acquisition features of the instrument. After testing diverse average scans numbers and resolution values, these parameters were set at 512 scans per spectra and 16 cm^{-1} respectively. Laboratory samples were analysed with the FT-NIRS instrument described for model's development, using the off-line arrangement, i.e. by submerging the probe into plastic tubes that contained the samples.

Figure 6, shows the changes in the spectral profile over the sequential increment of Biomass and Glycerol concentrations. The most relevant changes in this case are those related to the baseline shift generated by the presence of solid particles of Biomass, which is clearly remarked by the comparison with samples prepared in absence of this analyte. Additionally, the first and second O-H overtones bands (7000 and 5000 cm^{-1} respectively) show differences that can be attributed to changes in the intermolecular interactions of hydrogen bond that take place as soon as the Glycerol concentration increases in the solution.

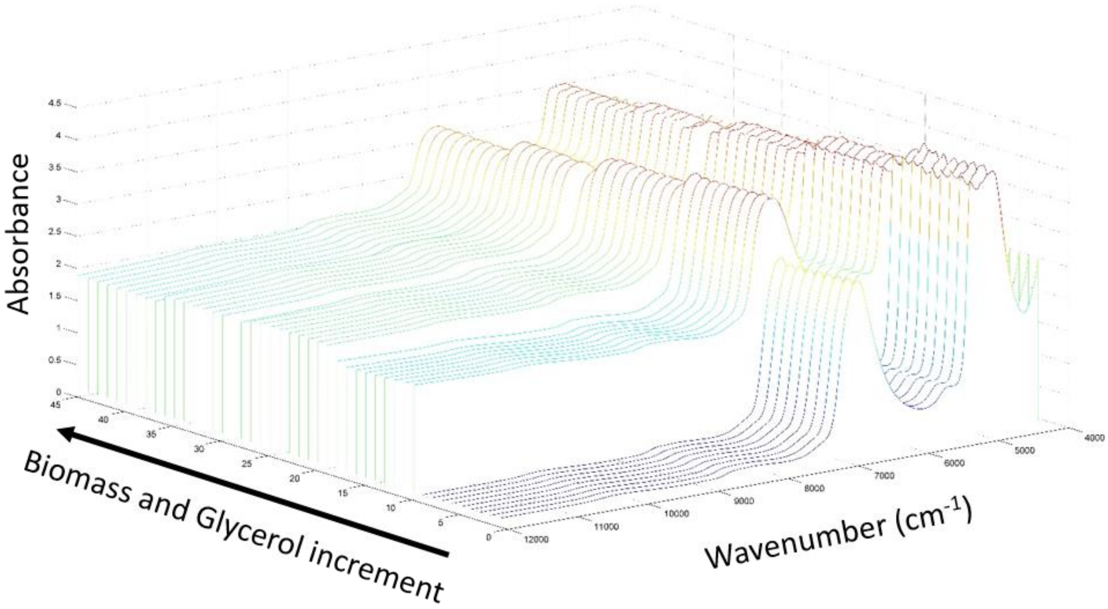


Figure 6. NIRS absorbance spectra of samples volumetrically prepared with Glycerol and Biomass stock solutions.

Spectral changes of Laboratory samples prepared with diverse concentrations of Total protein had a more random and unclear spectral trend than those observed in the samples prepared with Glycerol, possibly due to the range of concentrations studied, which was noticeable lower. However, for both sets of samples, preliminary PLS models were constructed, with the aim of evaluating their performance in the prediction of samples from the process. The lack of sources of variability in these initial regressions was evident during such evaluation, producing RMSEP tremendously high (data not shown).

4.4.2 Inclusion of data collected off-line

A total of eleven cultivations were run during this study. For all of them, reference values were determined based on a sampling procedure enabled by an outlet of the reactor especially prepared for such purpose (as indicated in Figure 1B), and following the methodologies described in section 1.2.2.2, NIRS measurements were acquired off-line for the first eight of these processes and in-line for the last three. In the case of off-line analyses, the NIRS data was recorded from the same test tubes taken for the monitoring of the process evolution using the reference methods. It was done once more by immersing the transflectance probe into the plastic tubes.

The NIRS data of the last three processes was recorded totally in-line, as will be explained in the next section. What is important to describe at this point is the general profile of the process variables studied, as well as their changes over the process evolution. According to the Glycerol addition strategy, the eleven cultivation processes can be classified in one pulse, two pulses and sequential pulses processes. Table 2 displays the different values of the carbon source incomes and Biomass production for all the cultivations considered.

Table 2. Classification of the cultivations according to the strategy employed for Glycerol addition.

cultivation ID	G-b (g/L)	G-fb (g/L)	B (g/L)	Strategy of Glycerol addition	NIRS data collection
FJ1702	35	55	45	Exponentially fed	Off-line
FJ1703 and FJ1705	35	145	90		
FJ1707	35	165	100		
FJ1706 and FJ1708	40	140	90		
FJ1711	40	245	140		
FJ1704 and FJ1713	82	90	50	Two pulses	Off-line In-line
FJ1715	40	140	60		In-line
FJ1714	75	4 pulses, range 5-25	68	Sequential pulses	In-line

ID: Identification Code; B: Maximum concentration of Biomass reached in the process; G-b: Glycerol added in the batch stage; G-fb: Glycerol added in the fed-batch stage.

Figure 7, shows an example of the evolution of each one of the three different Glycerol addition strategies studied, as a guide for the understanding of the general trends of the process variables. These graphics were created from reference values.

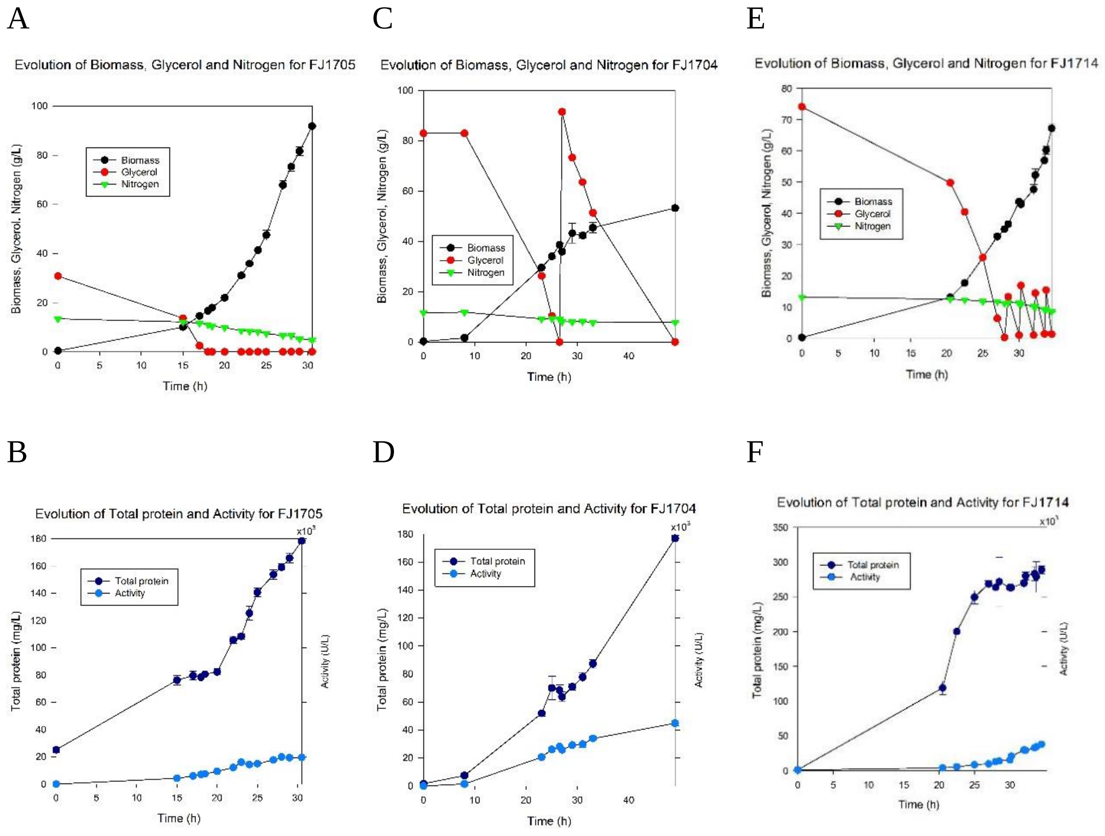


Figure 7. Examples of process variables evolution for strategies of *(**A**)* and *(**B**)* one pulse, (***C***) and (***D***) two pulses and (***E***) and (***F***) sequential pulses of Glycerol addition.

Due to the NIRS data acquired off-line was recorded from samples taken from the reactor for monitoring by reference methods, the number of spectra available for cultivations recorded off-line (from FJ1702 to FJ1711) is in the range 10-16 spectra per process. Figure 8, shows an example of the absorbance NIRS data obtained for a cultivation with two pulses of Glycerol (FJ1704). In this figure it is possible to observe the progressive baseline shift attributed to the increment of Biomass concentration in a succession with more changes than those observed during the study of the samples volumetrically prepared at the laboratory (Figure 6). By this fact, these series of spectra show the complexity of the chemical changes that take place during the cultivation process and which were not included in the first set of laboratory samples. Those variability sources are reflected mainly in the changes of intensity of the absorbance bands at 9000, 7000, 5800 and 5000 cm^{-1} (related to C-H, O-H and N-H vibrations).

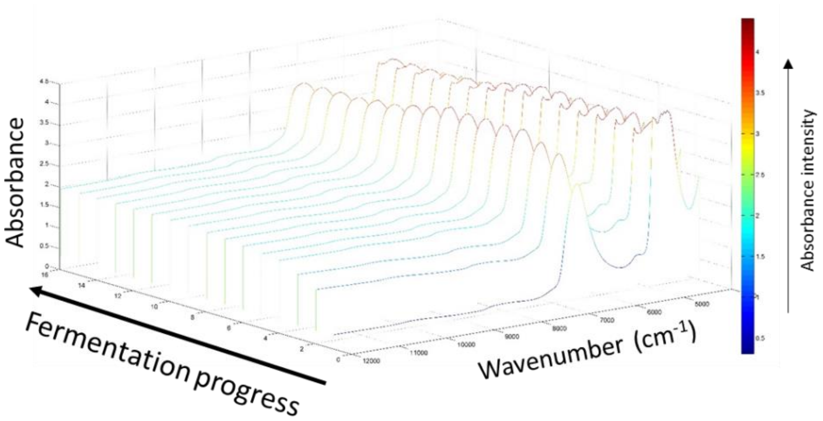


Figure 8. Example of the NIRS data acquired off-line from a process completed with Glycerol addition in two pulses of Glycerol (FJ1704).

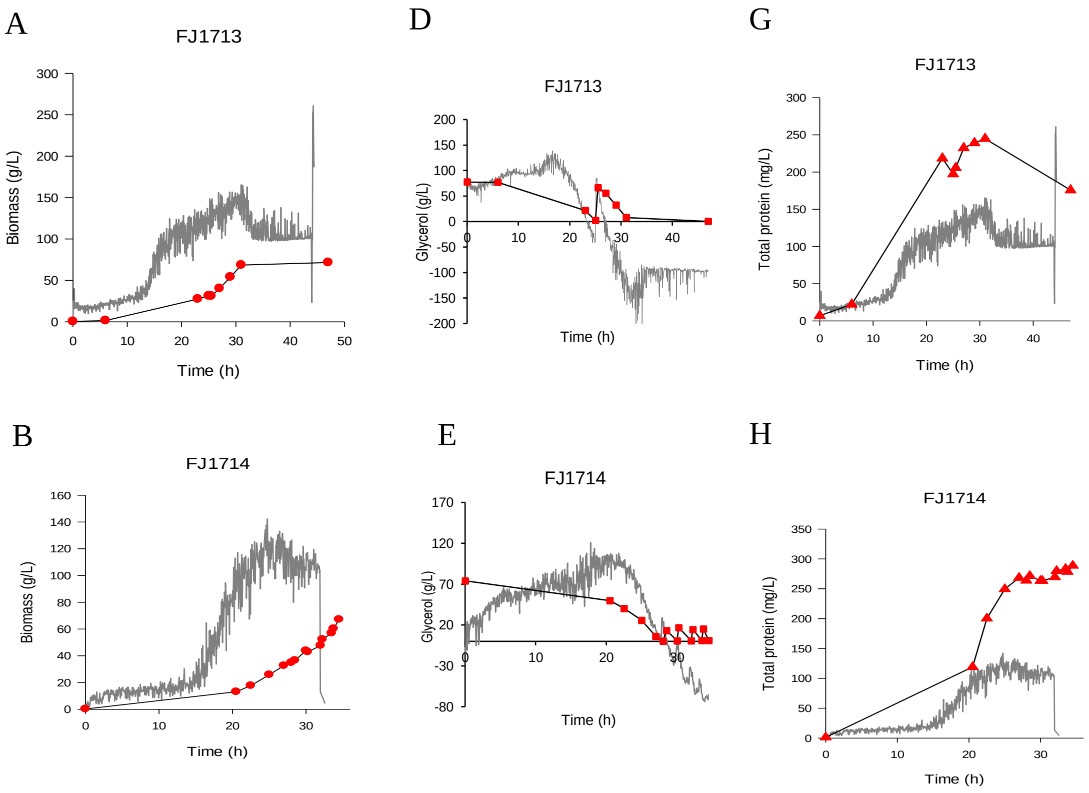
New PLS regressions were calculated by including into the models created with the samples volumetrically prepared, the NIRS data acquired off-line from samples produced during the first eight cultivations. For most of the analytes, results were far from the minimum proper performance of a good prediction model, even during internal assessment for calibration. Due to the goal of the study was developing an analytical method robust enough for the monitoring of cultivations following all the possible strategies of Glycerol addition and considering as much of sources of variability as possible, this labour was done considering data from bioprocess executed using both exponentially fedand two pulses of Glycerol addition strategies (samples from the sequential addition of Glycerol were not included because only one experience like this was run, and it was recorded in-line). With this objective in mind, samples prepared volumetrically were taken out of calibration sets of the first models created, and new models were developed for Glycerol and Total protein, based only on NIRS data acquired off-line. Only in one exercise of Biomass and Total protein modelling, it was possible to keep the laboratory samples data with the data acquired off-line in the same calibration set. The task of finding the appropriate data pre-treatment and spectral range was off course harder this time than in the first attempt. Particularly for Glycerol it was difficult to find a useful combination of parameters during modelling. Table 3, describes the characteristics of the models developed using the NIRS data acquired off-line.

Table 3. Figures of merit of PLS models developed based on NIRS data acquired off-line from cultivations run following the exponentially fed and two pulses of Glycerol addition strategies.

Analyte	LV	PT	SR (cm-^{1})	R^2	CR	RMSEC
Biomass (g/L)	5	SNV	11987-7791	0.96	5-100	6
Glycerol (g/L)	6	SNV	7413-5539	0.996	0-75	1
Total protein (mg/L)	2	OSC	9311-5639	0.995	0-300	5

LV: Latent Variables (PLS factors); PT: Data Pre-treatment; SNV: Standard Normal Variate; OSC: Orthogonal Signal Correction; SR: Spectral Range; R^2: Coefficient of determination for the PLS regression; CR: Concentration Range; RMSEC: Root Mean Square Error of Calibration. CR and RMSEC are expressed in the concentration units of the corresponding analyte.

The performance of these models was assessed by the prediction of NIRS data acquired in-line: two experiences following the two pulses strategy (FJ1713 and FJ1715) and one experience following the sequential addition of Glycerol during the fed-batch stage (FJ1714). The results of predictions are displayed in Figure 9. It can be observed that the general trend of the process profile outlined by the NIRS predictions was similar to those outlined by the reference values. However, very important bias which still had to be overcome. Solving such bias error was the objective of the next part of the work.



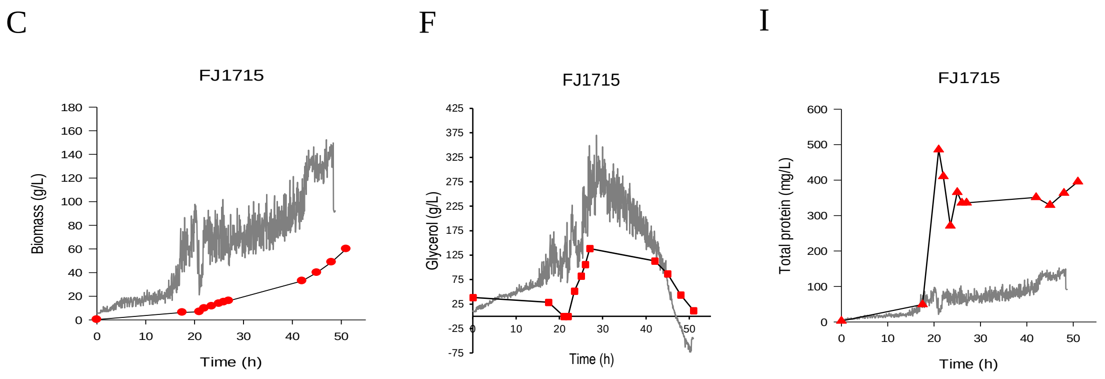


Figure 9. Biomass, Glycerol and Total protein evolution over processes monitored in-line. Values predicted by models calculated with NIRS data acquired off-line are plotted in grey, and red points represent reference data. Figures (***A***), (***D***), (***G***), (***C***), (***F***) and (***I***) show predictions of cultivations run with the strategy of addition of two pulses of Glycerol. Figures (***B***), (***E***) and (***H***) display results for the strategy of sequential addition of 4 pulses of Glycerol during the fed-batch stage.

4.4.3 Inclusion of data collected in-line

The NIRS data acquired in-line, by the immersion of the probe directly into the bioprocessing reactor, compress the biggest complexity and number of spectra of this work. Its inclusion in the calibration sets was the last and most critical step for the optimization of the models. For the three processes recorded under this condition, the transflectance probe was immersed in the reactor from the very beginning of the cultivation. Indeed, it was included in the system before the sterilization process, previous of course to the inoculation. The total amount of spectra employed for modelling from the in-line measurements was 875 for FJ1713, 639 for FJ1714 and 977 for FJ1715. The time between each acquisition was set at 3 minutes for all the three experiments, by means of the advanced acquisition window of the Opus software. Therefore, differences in the final amounts of spectra are related to the total of hours that each process required to be completed. Figure 10, presents an example of the absorbance spectra acquired during the cultivation run following the strategy of sequential addition of Glycerol during the fed-batch stage (FJ1714).

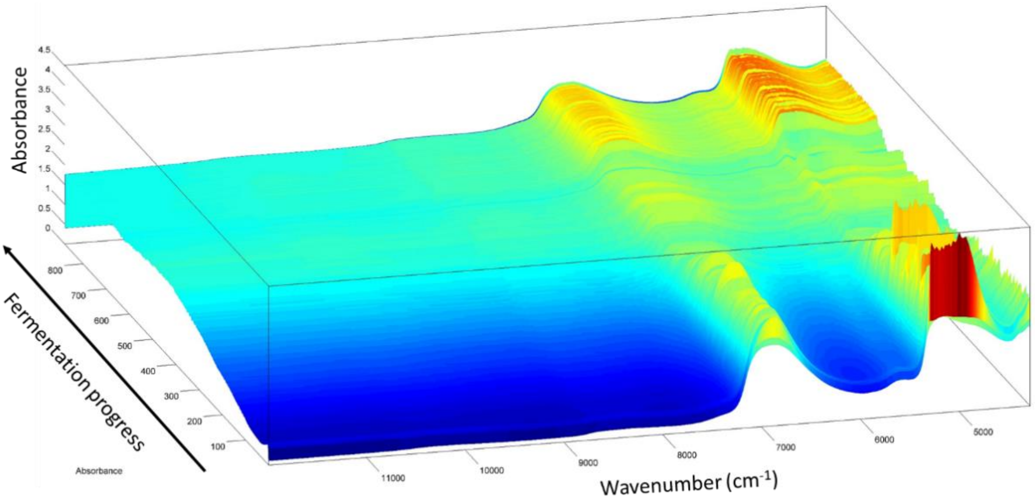


Figure 10. Absorbance NIRS data acquired in-line every 3 min, during the cultivation of sequential addition of Glycerol during the fed-batch stage.

An exploratory analysis of the NIRS data acquired in-line confirmed that Biomass concentration was the main source of variability of the process. Figure 11, displays a scores plot of a PCA of the NIRS data of FJ1714 after SNV pre-treatment. Changes along the first component axe (which explains the 95.49% of the variability of the system) follow the progressive increment of Biomass concentration observed in this cultivation.

On the other hand, samples are also widely distributed along the projection of the second component of the scores plot presented. This fact indicates that other complex chemical changes (as simultaneous variations on substrate and metabolites concentrations) as well as physical changes (bubbles, foam, mechanical agitation, increment of solid particles in the medium) produce variability that can definitely be detected by NIRS. The generous amount of data provides detailed evidence of NIR spectral changes during the process.

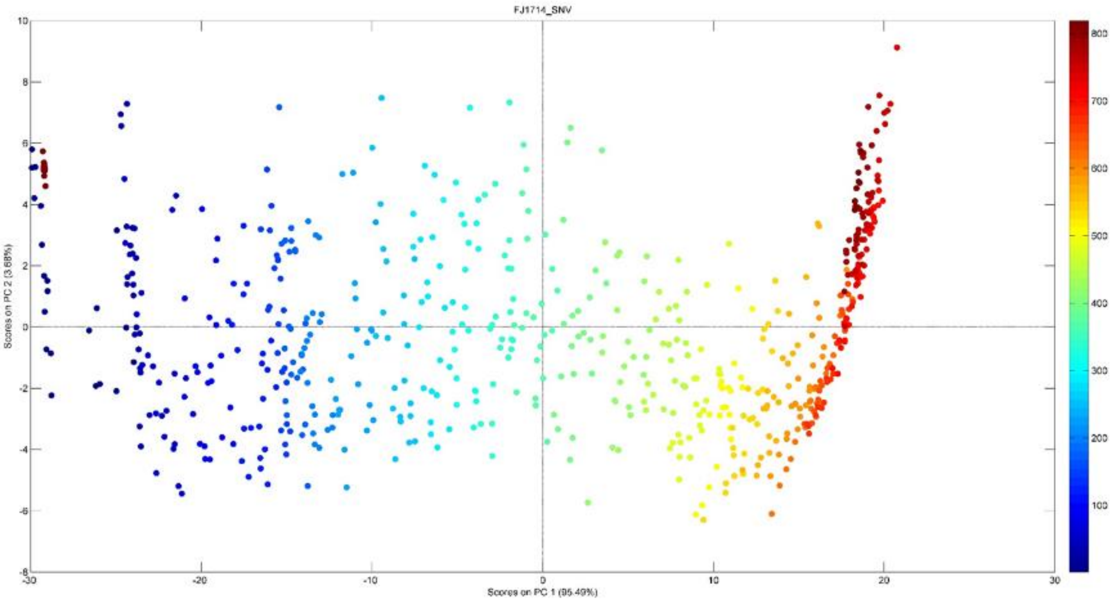


Figure 11. Scores plot of FJ1714 after SNV data pre-treatment.

The data recorded using the in-line approach was initially included into the models developed using the off-line data collection methodology. The assignation of reference values to NIR individual spectra was completed using the sampling times and the NIRS acquisition times, considering the inoculation as time 0 h. Once more, the lack of variability in the methods developed using only off-line data was reflected in modelling difficulties using simultaneously data collected both off-line and in-line. PLS regressions shown two trends, clearly different between the two kinds of data collection: off-line and in-line. Keeping in mind that the goal of the study was the development of methods robust enough for the in-line monitoring of *P. pastoris* cultivations, data collected off-line was excluded of the calibration sets. Afterwards, new models considering only data collected in-line, were constructed. The accuracy of such models was evaluated as the root mean standard error of prediction (RMSEP) for the points with reference values available (internal validation). As external validation feature, the standard deviation (SD) of the predictions of spectra without reference values -which were not included into the calibration set- was employed as a criterion for the selection of the final models. Such SD results were obtained multiplying by 2 the deviation value generated per spectrum by the Unscrambler 10.3, after challenging the models using the Prediction task.

The first attempt done considering only data collected in-line in the calibration sets was done considering all the three data sets together. Table 4, summarizes the results for this experience in the fed batch stage of the cultivations (after 20.5h). The small amount of reference values available during the batch stage was a factor that made hard finding appropriate calibration features during this section of the process. Since the amount of reference values available from the batch stage was the same for all the cultivations studied (only one point), results have the same trend of poor predictive capability in this period for all the models. Because of this fact, all the errors of prediction displayed in this work are calculated using values after 20.5 h. Figure 12, illustrates these results for the bioprocess with sequential addition of glycerol.

Table 4. Figures of merit of models created considering the two kind of Glycerol addition strategies together.

Analyte	**LV**	**PT**	**SR (cm-1)**	**CR**	**RMSEC**	**RMSEP A**	**SD *A***	**RMSEP *B***	**SD *B***	**SEL**
Biomass (g/L)	3	2D	7590-6078	5-75	3	33	10-20	8	7-17	0.8
Total protein (mg/L)	6	1D	8925-5515	7-485	11	16	46-120	9	48-102	11
Glycerol (g/L)	5	RN	7413-5492	2-138	7	69	40-80	15	10-20	-
Nitrogen (g/L)	4	2D+ SNV	11987-4296	6-15	1	1	2-4	1	3-7	0.27
Activity (U/L)	7	None	11987-4296	3000-45000	2100	2919	5000-7500	2000	5000-10000	580

LV: Latent Variables (PLS factors); PT: Data Pre-treatment; 1D: First Derivative Savitzky-Golay; 2D: Second Derivative Savitzky-Golay; RN: Range Normalization; SNV: Standard Normal Variate; SR: Spectral range; CR: Concentration Range; RMSEC: Root Mean Square Error of Calibration; RMSEP: Root Mean Square Error of Prediction SD: Standard Deviation; ***A****: set of cultivations with addition of glycerol in two pulses;* ***B****: cultivation with sequential addition of glycerol; SEL: Standard Error of Laboratory (SD of reference methods). CR, RMSEC, RMSEP as well as SD are expressed in the corresponding units per analyte.*

Figure 12, also shows that the bias error previously found between the values predicted for NIRS data collected in-line and the reference values, decreased in an important ratio by the creation of calibration sets considering only data collected in-line. This can be understood considering the sources of variability that are present during the process with relevant impact into the NIRS data. Some of these sources are the 30°C of temperature (higher than the temperature for off-line measurements), the eventual formation of foam, the permanent and intense agitation inside of the reactor (500-1000 rpm) as well as the aeration of the system (between 0-1 vvm). All these factors have a proved effect on the NIRS data, as has been previously demonstrated [8].

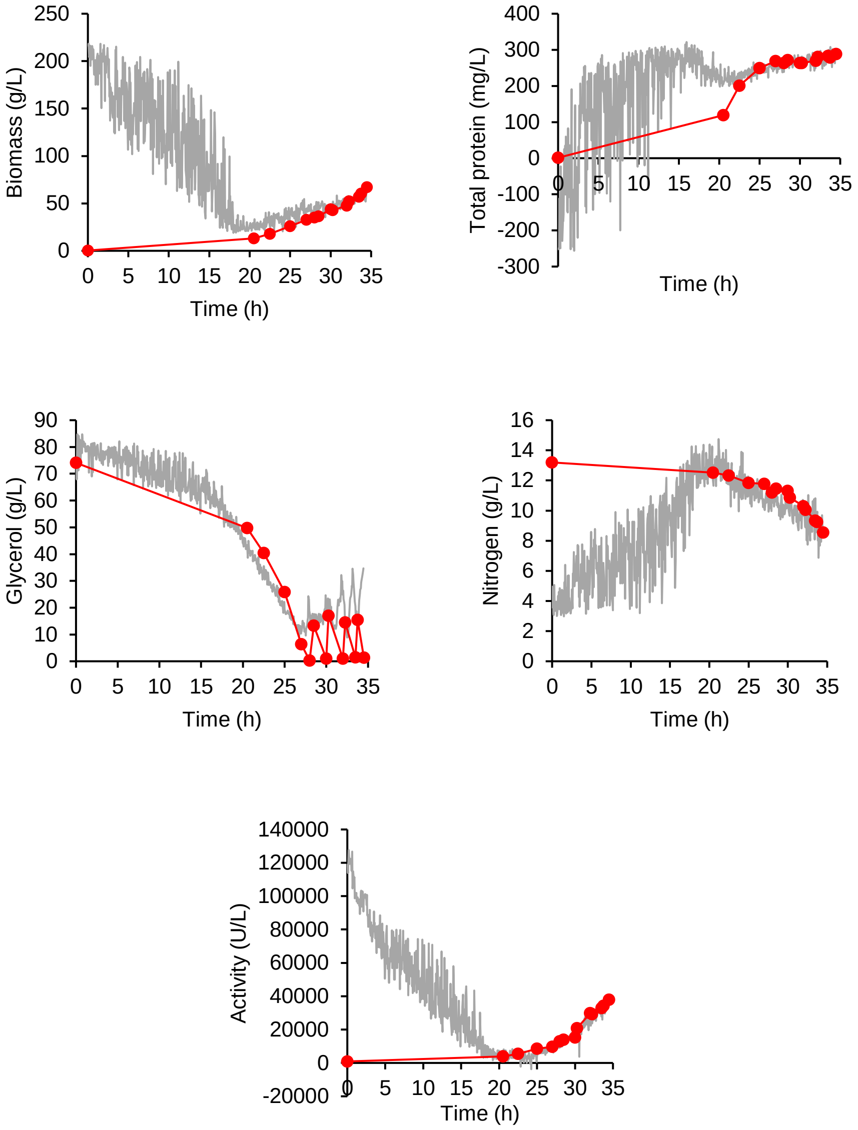


Figure 12. Biomass, Glycerol, Total protein, Nitrogen and Lipolitic activity evolution over bioprocess completed using the sequential addition of Glycerol (FJ1714), using models described in Table 4 (calculated using data from cultivations with different strategies of Glycerol addition during the calibration). Values predicted from NIRS data are plotted in grey, and red points represent reference data.

Nevertheless, high SD values were found for predictions of models constructed using together all the data collected in-line (mixing different Glycerol addition strategies). To study the effect of the Glycerol addition strategy on modelling the NIRS data, new models were calculated considering two different data sets, generated from splitting the available data sets regarding the strategies of Glycerol addition employed. Table 5, presents the results for these final models.

The main improvement reached using the split data was the simplification of the models, which was evidenced by the reduction of the number of PLS factors required for calibrations and predictions reported in Table 5, compared to those in Table 4. This aspect is particularly interesting in the case of the Biomass models. The number of PLS factors required for the model constructed using the in-line data from bioprocess with the two strategies of Glycerol addition together, was 3. However, when models for Biomass were calculated with the split data, the number of factors required for the model created with data from cultivations where Glycerol was added in two pulses (FJ1713 and FJ1715), increased to 6, while decreased to 2 for the model calculated with the data from the sequential addition of Glycerol during the fed-batch stage (FJ1714). These values allow to deduce that the complexity of the NIRS data from cultivations where the Glycerol is added only in two pulses is higher to the complexity of data from processes where the addition of Glycerol is done sequentially.

Additionally, for Total protein, Nitrogen and Lipolitic activity, the ranges of the SD calculated based on the prediction of samples acquired in-line and excluded during calibration, were also lower than those obtained using the two strategies of Glycerol addition together.

Table 5. Figures of merit of models calculated using the data collected in-line split according to the strategy of Glycerol addition.

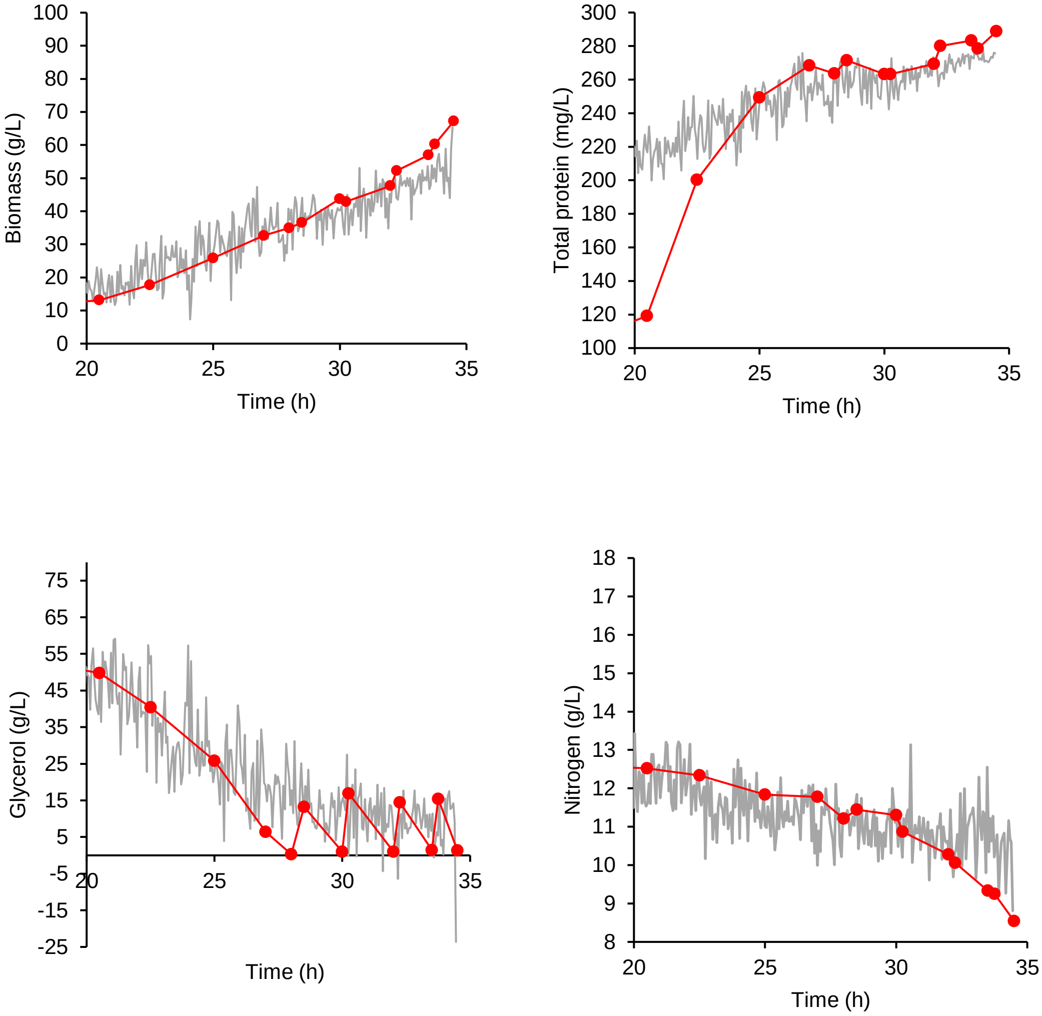
	Biomass	**Total protein**	**Glycerol**	**Nitrogen**	**Lipolitic activity**
Calibration set	*Data from FJ1713 and FJ1715 correlated to reference values*				
Validation set	*All the data from FJ1713 excluded during modelling*				
SEL	0.8 g/L	11 mg/L	-	0.27 g/L	580 U/L
LV	6	2	5	2	2
Data pre-treatment	2D	1D	2D+SNV	2D+SNV	None
Spectral range (cm^{-1})	7590-7035	9320-5415	11987-4296	11987-4296	11987-4296

	Biomass	Total protein	Glycerol	Nitrogen	Lipolitic activity
Range of prediction	5-75 g/L	119-289 mg/L	10-50 g/L	8-14 g/L	13000-45000 U/L
RMSEC	3 g/L	14 mg/L	6 g/L	0.3 g/L	781 U/L
RMSEP	15 g/L	5 mg/L	7 g/L	0.7 g/L	1553 U/L
SD of NIRS prediction	20-50 g/L	25-27 mg/L	80-120 g/L	1.5-2.5 g/L	5000-75000 U/L
Calibration set	*Data from FJ1714 correlated to reference values*				
Validation set	*All the data from FJ1714 excluded during modelling*				
LV	2	2	2	3	3
Data pre-treatment	2D	1D	2D+SNV	2D+SNV	None
Spectral range (cm^{-1})	7845-6078	8925-5515	8007-6502	11987-4296	11987-4296
Range of prediction	5-75 g/L	119-289 mg/L	10-50 g/L	8-14 g/L	13000-45000 U/L
RMSEC	3 g/L	1 mg/L	2 g/L	0.15 g/L	1517 U/L
RMSEP	4 g/L	5 mg/L	2 g/L	0.8 g/L	3010 U/L
SD of NIRS prediction	10-30 g/L	7-22 mg/L	20-40	1.5-4 g/L	3045-10105 U/L

LV: Latent Variables (PLS factors); SEL: Standard Error of Laboratory (average of Standard Deviation of reference values); 1D: First Derivative Savitzky-Golay; 2D: Second Derivative Savitzky-Golay; SNV: Standard Normal Variate; SD: Standard Deviation; RMSEC: Root Mean Square Error of Calibration; RMSEP: Root Mean Square Error of Prediction; SD of NIRS: SD calculated multiplying by 2, the values of deviation from the task of prediction of Unscrambler 10.3.

Figure 13, illustrates the predictive capability of models calculated using in-line data from the cultivation with the strategy of addition of Glycerol by sequence of pulses (FJ1714), for the prediction of those spectra excluded during the calibration. These models achieved the better performance in the study presented. However, it is clear that further external validations with new data obtained using the same collecting conditions as well as the same cultivation conditions and Glycerol adding strategy, are necessary to evaluate the robustness of these models. Sampling for generating more reference values during the batch stage, would also provide useful data for enhancing the predictive capability of these models.

In the particular case of Glycerol, the error in the prediction was still too high regarding the concentration range that was intended to be controlled. Because of that, this model is not considered useful for the in-line process monitoring approach. The models calculated for the Biomass, Total protein and Nitrogen concentration, and Lipolitic Activity, can be used as a tool for monitoring the general trend of the process, always bearing in mind the SD found for each of them.



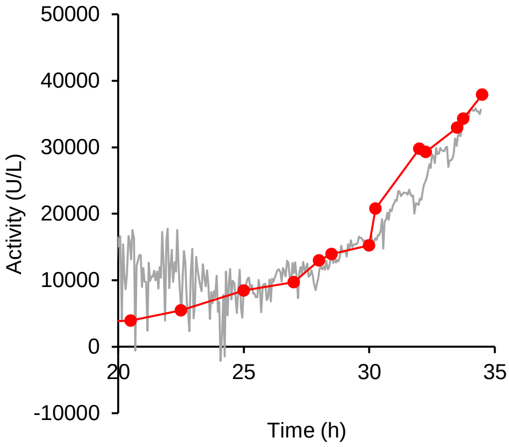


Figure 13. Biomass, Glycerol, Total protein, Nitrogen and Lipolitic activity evolution over cultivation completed using the sequential addition of Glycerol (FJ1714), using models described in Table 5 (using only data from FJ1714). Values predicted from NIRS data are plotted in grey, and reference data is represented by red points.

4.5 Conclusions

Transflectance is the acquisition mode that enables the simpler instrumental conditions for in-line monitoring of the recombinant production of Lipase B from *C. anctartica* in *P. pastoris* using Glycerol as carbon source. Using a FT-NIR spectrometer with fiber optic transflectance probe, it was possible to develop models for the in-line monitoring over the fed-batch stage of Biomass, Total protein, Nitrogen and Lipolitic Activity in the ranges of interest. The models calculated for Glycerol produced prediction values with a standard deviation higher than the concentration range of interest. The most important source of variability for NIRS data collected from the mentioned bioprocess system is the Biomass concentration. However, agitation, foaming, temperature and aeration are also sources of variability that can impact the predictive capability of the models. Consequently, the development of more accurate methods based on NIRS requires as much data collected in-line as possible, as well as an increased number of reference values over all the stages of the process (both from the batch and the fed-batch stages). Finally, it was found that the predictive capability of the models is affected by the strategy employed for the addition of the carbon source to the system.

4.6 References

[1] O. Kirk and M. W. Christensen, “Lipases from Candida antarctica: Unique biocatalysts from a unique origin,” *Org. Process Res. Dev.*, vol. 6, no. 4, pp. 446–451, 2002.

[2] J. M. Robert *et al.*, “Production of recombinant lipase B from Candida antarctica in Pichia pastoris under control of the promoter PGK using crude glycerol from biodiesel production as carbon source,” *Biochem. Eng. J.*, vol. 118, pp. 123–131, 2017.

[3] J. L. Cereghino and J. M. Cregg, “Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*,” *FEMS Microbiol. Rev.*, vol. 24, no. 1, pp. 45–66, Jan. 2000.

[4] J. Crowley, S. A. Arnold, N. Wood, L. M. Harvey, and B. Mcneil, “Monitoring a high cell density recombinant Pichia pastoris fed-batch bioprocess using transmission and reflectance near infrared spectroscopy,” *Enzyme Microb. Technol.*, vol. 36, pp. 621–628, 2005.

[5] G. P. Lin-Cereghino, J. Lin-Cereghino, C. Ilgen, and J. M. Cregg, “Production of recombinant proteins in fermenter cultures of the yeast Pichia pastoris,” *Curr Opin Biotech*, vol. 13, no. 4, pp. 329–332, 2002.

[6] M. L. Fazenda *et al.*, “Towards better understanding of an industrial cell factory,” *Microb. Cell Fact.*, vol. 12, pp. 1–14, 2013.

[7] K. Kiviharju, K. Salonen, U. Moilanen, and T. Eerikäinen, “Biomass measurement online: the performance of in situ measurements and software sensors,” *J Ind Microbiol Biotechnol*, vol. 35, pp. 657–665, 2008.

[8] E. Tamburini, M. Marchetti, and P. Pedrini, “Monitoring Key Parameters in Bioprocesses Using Near-Infrared Technology,” *Sensors*, vol. 14, no. 10, pp. 18941–18959, Oct. 2014.

[9] S. Beutel and S. Henkel, “In situ sensor techniques in modern bioprocess monitoring,” *Appl. Microbiol. Biotechnol.*, vol. 91, no. 6, pp. 1493–1505, 2011.

[10] A. Mitic *et al.*, “Implementation of Near-Infrared Spectroscopy for In-Line Monitoring of a Dehydration Reaction in a Tubular Laminar Reactor,” *Org. Process Res. Dev.*, vol. 20, no. 2, pp. 395–402, Feb. 2016.

[11] A. E. Cervera, N. Petersen, A. E. Lantz, A. Larsen, and K. V. Gernaey, “Application of near-infrared spectroscopy for monitoring and control of cell culture and cultivation,” *Biotechnol. Prog.*, vol. 25, no. 6, pp. 1561–1581, 2009.

[12] A. Trilli, E. Tamburini, G. Vaccari, and S. Tosi, "Near-Infrared Spectroscopy: A Tool for Monitoring Submerged cultivation Processes Using an Immersion Optical-Fiber Probe," *Appl. Spectrosc.*, vol. 57, no. 2, pp. 132–138, 2003.

[13] M. Goldfeld *et al.*, "Advanced near-infrared monitor for stable real-time measurement and control of Pichia pastoris bioprocesses," *Biotechnol. Prog.*, vol. 30, no. 3, pp. 749–759, 2014.

[14] S. Kim *et al.*, "Real-time monitoring of glycerol and methanol to enhance antibody production in industrial Pichia pastoris bioprocesses," *Biochem. Eng. J.*, vol. 94, pp. 115–124, 2015.

[15] P. Biechele, C. Busse, D. Solle, T. Scheper, and K. Reardon, "Sensor systems for bioprocess monitoring," *Eng. Life Sci.*, vol. 15, no. 5, pp. 469–488, 2015.

[16] X. Garcia-Ortega, N. Adelantado, P. Ferrer, J. L. Montesinos, and F. Valero, "A step forward to improve recombinant protein production in Pichia pastoris: From specific growth rate effect on protein secretion to carbon-starving conditions as advanced strategy," *Process Biochem.*, vol. 51, no. 6, pp. 681–691, 2016.

[17] M. M. Bradford, "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding," *Anal. Biochem.*, vol. 72, no. 1–2, pp. 248–254, May 1976.

[18] A. Tabacco, F. Meiattini, E. Moda, and P. Tarli, "Simplified Enzymic / Colorimetrlc Serum Urea Nitrogen Determination Thymol Is a Suitable Preservative for Uric Acid Standards in the Uricase Technique CHEMISTRY , in the CK-BB Least-Squares Evaluation of Linearity More on the Detection of Serum CK-BB Acti," *Clin. Chem.*, vol. 25, no. 2, pp. 4–5, 1979.

[19] J. K. Fawcett and J. E. Scott, "A rapid and precise method for the determination of urea.," *J. Clin. Pathol.*, vol. 13, no. 2, pp. 156–9, Mar. 1960.

[20] D. M. Freire, E. M. F. Teles, E. P. S. Bon, and G. L. S. Anna, "Lipase production by Penicillium restrictum in a bench-scale fermenter," *Appl. Biochem. Biotechnol.*, vol. 63–65, no. 1, pp. 409–421, Mar. 1997.

[21] A. Paul, P. Carl, F. Westad, J.-P. Voss, and M. Maiwald, "Towards Process Spectroscopy in Complex cultivation Samples and Mixtures," *Chemie Ing. Tech.*,

vol. 88, no. 6, pp. 756–763, Jun. 2016.

[22] S. Alison Arnold, L. Matheson, L. M. Harvey, and B. Mcneil, “Temporally segmented modelling: A route to improved bioprocess monitoring using near infrared spectroscopy?,” *Biotechnol. Lett.*, vol. 23, no. 2, pp. 143–147, 2001.

[23] B. Finn, L. M. Harvey, B. McNeil, and B. McNeil, “Near-infrared spectroscopic monitoring of biomass, glucose, ethanol and protein content in a high cell density baker’s yeast fed-batch bioprocess,” *Yeast*, vol. 23, no. 7, pp. 507–517, 2006.

[24] M. Alcalà *et al.*, “Near-infrared Spectroscopy in Laboratory and Process Analysis,” *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, pp. 1–39, 2012.

4. *Annex 1*

A. Y. Miró Vera and M. Alcalà Bernàrdez, “Near-Infrared Spectroscopy in Identification of Pharmaceutical Raw Materials,” *Encyclopedia of Analytical Chemistry*. John Wiley & Sons, Ltd, pp. 1–19, 2017.