IOWA STATE UNIVERSITY
Digital Repository

Graduate Theses and Dissertations | Graduate College

2011

Single seed discriminative applications using near infrared technologies

Lidia Esteve Agelet
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd

 Part of the Bioresource and Agricultural Engineering Commons

Recommended Citation
Esteve Agelet, Lidia, "Single seed discriminative applications using near infrared technologies" (2011). *Graduate Theses and Dissertations*. 12023.
http://lib.dr.iastate.edu/etd/12023

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Single seed discriminative applications using near infrared technologies

by

Lidia Esteve Agelet

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Agricultural and Biosystems Engineering

Program of Study Committee:
Charles R. Hurburgh Jr., Major Professor
Carl J. Bern
Brian L. Steward
Dianne H. Cook
Kenneth J. Moore

Iowa State University

Ames, Iowa

2011

Copyright© Lidia Esteve Agelet, 2011. All rights reserved.

TABLE OF CONTENTS

ABSTRACT ..... **V**

ACKNOWLEDGEMENTS ..... **VI**

CHAPTER 1. GENERAL INTRODUCTION **1**

RATIONALE 1

DOCUMENT ORGANIZATION 2

LITERATURE REVIEW 2

1. Theory behind Near Infrared Spectroscopy *2*

1.1 Infrared in the Electromagnetic Spectrum 2

1.2 Theory behind NIR Absorption 4

1.3 Using Light in Spectroscopy 8

2. Instrumentation: Spectrometers *13*

2.1. Sample Compartment 13

2.2. Light Source 14

2.3. Wavelength Selection 15

2.4. Detectors 18

3. Other Near Infrared Technologies *20*

3.1. Fourier Transform NIR (FT-NIR) 20

3.2. Hyperspectral Imaging 21

4. Chemometrics *24*

4.1. Selecting Samples, Reference Methods, and Spectral Data 26

4.2. Spectra Pretreatments 28

4.3. Principal Component Analysis (PCA) 33

4.4. Linear Calibration Models 36

4.5. Non-linear Models 40

4.6. Pattern Recognition and Classification 55

4.7. Variable Selection 61

4.8. Validation Procedure and Statistics 63

5. Near Infrared Spectroscopy for Single Seed Analyses *65*

5.1 NIRS applications in quantitative analysis of single seeds..... 65
5.2 NIRS for seed sorting and discrimination..... 76
JUSTIFICATION FOR WORK AND OBJECTIVES..... 84
REFERENCES..... 85

CHAPTER 2. NEAR INFRARED REFLECTANCE SPECTROSCOPY APPLIED TO DISCRIMINATION OF CONVENTIONAL AND ROUNDUP READY SOYBEAN SEEDS..... 103

ABSTRACT..... 103
INTRODUCTION..... 104
EXPERIMENTAL..... 106
RESULTS AND DISCUSSION..... 115
CONCLUSIONS..... 121
REFERENCES..... 123

CHAPTER 3. DISCRIMINATION OF CONVENTIONAL AND ROUNDUP READY SOYBEAN SEEDS. TRANSMITTANCE VS REFLECTANCE MEASUREMENTS AND MOISTURE EFFECT..... 124

ABSTRACT..... 124
INTRODUCTION..... 125
MATERIAL AND METHODS..... 128
RESULTS AND DISCUSSION..... 135
CONCLUSIONS..... 146
REFERENCES..... 148

CHAPTER 4. FEASIBILITY OF NEAR INFRARED SPECTROSCOPY FOR ANALYZING CORN KERNEL DAMAGE AND VIABILITY OF SOYBEAN AND CORN KERNELS..... 150

ABSTRACT..... 150
1.INTRODUCTION..... 151
2.EXPERIMENTAL..... 155
3.RESULTS AND DISCUSSION..... 162

4.CONCLUSIONS..... 168
5.REFERENCES 169

CHAPTER 5. TRAINING ON NEAR INFRARED TECHNOLOGIES 172

ABSTRACT 172
INTRODUCTION 173
ANALYSIS PHASE: APPROACHING YOUNG STUDENTS TO NIRS AND TO LABORATORY DYNAMICS 175
DESIGNING MATERIAL AND CHOOSING MEDIA..... 184
LEARNER PARTICIPATION AND OVERALL EVALUATION..... 195
CONCLUSIONS..... 196
REFERENCES 196

CHAPTER 6.GENERAL CONCLUSIONS..... 201

APPENDIX 1. LITERATURE REVIEW PAPER 203

APPENDIX 2. TRAINING MANUAL OF THE GRAIN QUALITY LABORATORY..... 218

ABSTRACT

Near infrared spectroscopy (NIRS) have been utilized in a wide selection of single seed applications because it provides fast and non-destructive measurements. Despite the limitation of small seed sizes, NIRS has led to successful results. In this dissertation we explored the feasibility of NIRS for several discriminative applications for corn and soybean seeds. The first application focused on discrimination of conventional and genetically modified Roundup Ready® soybeans. Classification accuracies ranged from 75 to 99% percent. The highest accuracies were obtained with a light tube instrument and with locally weighted principal component regression (LW-PCR) models with few samles represented. Artificial Neural Network (ANN) and Support Vector Machines models gave simmilar accuracies. The technologies performing worse were the low ressolution chemical imaging unit and the Fourier Transform transmittance instrument due to their sensitivity to seed positioning. Discrimination within a single variety was possible above 95% accuracies for most of the varieties. Moisture was proven to impact the classification due to interactions between water and carbohydrates (fiber). For this reason, this application would be feasible for breeders working in controlled seed moistures. Other applications such as discrimination of damaged corn kernels (heat and frost damage) and viability of corn and soybeans with NIRS were analyzed. Only discrimination of heat-damaged corn kernels was successful (accuracies above 95% using partial least squares discriminant analysis, PLS-DA); frost-damaged kernels and non-viable seeds could not be discriminated with any of the tested algorithms. This indicates that NIRS only detects changes in seeds due to damage and there is no relationship with its viability. The final remaining question is what the extent of damage that a seed may suffer to be detected by NIRS would be.

ACKNOWLEDGEMENTS

Especial thanks to my friends and great researchers Aoife Gowen and Carlos Esquerre for their support through my internship at University College Dublin, as well as Professor Colm O'Donell for hosting me. I also want to acknowledge Robert Cogdill for his help and advices at the early stages of my PhD. I highly appreciate the valuable contribution of Paul Armstrong and Jasper Tallada for sharing his instrument and his help in the entire set-up. I thank the continuous support of Buchi personnel Gael Rolland, Brad Miller, and Ron Rubinovitz, which has been outstanding through this research.

The great support of each member of my family has been highly valuable and irreplaceable at all times. I want to thanks my mother Maria Rosa, my father Francisco, and my brother Victor for their prayers and support all the times, their company through this entire journey despite the distance. I want to especially acknowledge my grandparents, who may God have them resting in peace, for raising me with love and good values during my childhood. I hope I can make them proud. Thanks to my dear grandmother Laura and aunty Laura, whose emails always cheered me up and gave me strength, for their unconditional faith on me. My gratitude goes to Jerry Pierce, my American father, for bringing the best craziness to my life those last years and for his living example of citizenship. I also want to dedicate the fruit of my effort to my dear friends both in US and Spain, who never forgot me and cheered me up. For being real friends and taking me out for a coffee in the toughest times, make me laugh, and remind me about the best things in life.

Finally, I want to acknowledge Charlie, Connie, Glen, Howard, and my officemates Nanning and Gretchen their support not only as colleagues, but also as friends. For both stressful and good times we shared, which brought us closer and made of the grain quality lab a second family to me.

CHAPTER 1. GENERAL INTRODUCTION

RATIONALE

Since near infrared light (NIR) was first utilized for analytical purposes back in the 1960s, this technology has experienced an impressive growth. The advantages of near infrared spectroscopy are well recognized; to name some, the non-destructive nature of the analysis, high speed, no sample preparation required, affordable and flexible instrumentation, and good precision when calibrations are well developed. For these reasons, more scientists and companies from diverse fields are becoming interested in learning about near infrared spectroscopy (NIRS) and calibration development. Maintaining congruency in terminology over such a multidisciplinary field is becoming a challenge.

The agriculture sector has been benefiting from NIR analysis for a long time –the first published application involved moisture measurement in seeds-. The acceptance by the American Cereal Chemist (ACC) and American Oil Chemists (AOC) associations of NIR spectroscopy for routine analysis of oil, protein, moisture, starch, of bulk grain and oil seeds dates from the 1980s. Common NIR bulk sample analyzers provide the average of 250g of kernels, which is good for average measurements of whole batches and often enough for farmers and industrial processes. However, breeders and specific applications may need to target individual seeds with certain characteristics or higher batch homogeneity. Near infrared technologies offer a valuable tool for measuring whole batches fast and efficiently. In this dissertation, the feasibility of several applications for seed discrimination involved in food safety, quality, and breeder's purposes is analyzed. The discrimination of roundup ready soybean seeds from conventional, damage on single corn kernels, and corn and soybean seed viability is analyzed using several NIR technologies and algorithms.

DOCUMENT ORGANIZATION

This dissertation is organized in an exhaustive literature review and four main sections or papers. The main part of the literature review introduces the basics of near infrared spectroscopy, instrumentation, and its calibration. Part of it has been modified and published in the journal of critical reviews in analytical chemistry. It also has been used to update the training manual of the laboratory. Both documents can be found in the appendices section. The second main section reviews current research on near infrared technologies applied to single seeds. It will be prepared as a review manuscript and submitted to the Cereal Science journal. Three of the papers of the dissertation are intended to be submitted in four different recognized journals in either near infrared spectroscopy or cereal science fields. The fourth paper, which focuses on training in NIRS in the Grain Quality Laboratory, has been submitted to the journal of technology studies and if not accepted, may be modified and submitted to the journal of near infrared spectroscopy as an analysis of the problems when teaching about NIRS.

LITERATURE REVIEW

1. Theory behind Near Infrared Spectroscopy

1.1 Infrared in the Electromagnetic Spectrum

Light radiation is electromagnetic energy which can be arranged as an electromagnetic spectrum (Figure 1) according to properties such as wavelength, frequency, polarity, and intensity. Radiation energy is indirectly proportional to its wavelength, but directly proportional to the frequency. According to this, gamma rays in Figure 1 are the most energetic, while radio waves cover the longest wavelengths, shortest frequencies, and are the least energetic.

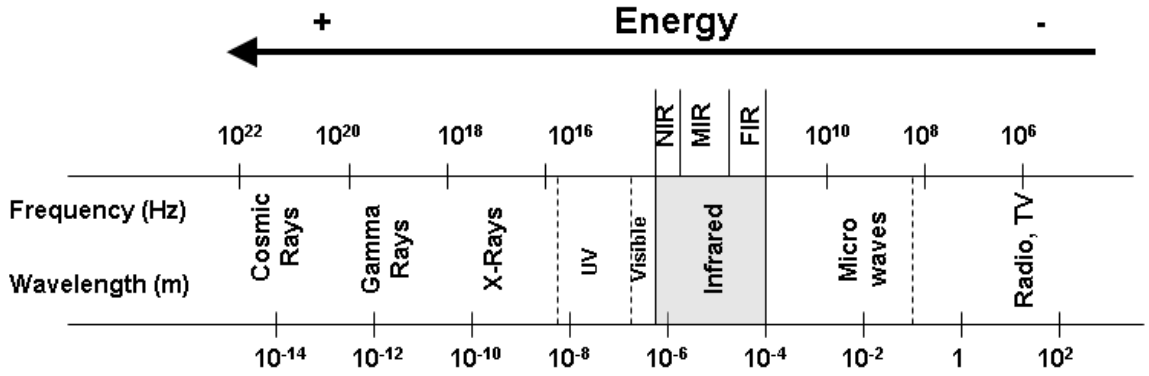


Figure 1. Electromagnetic spectrum

The infrared region is approximately located in the middle of the electromagnetic spectrum. It covers the electromagnetic wave frequencies in the range of 300 GHz to 400 THz (or wavelengths ranging from 1 mm to 750 nm in wavelength units) There are three differentiated regions in the infrared range: Far infrared (300 GHz (1 mm) to 30 THz (10 μm)), mid infrared (30 to 120 THz or 10 to 2.5 μm), and near infrared (from 120 to 400 THz or 2,500 to 750 nm).

The least energetic far infrared region (FIR) is utilized in the emerging technology of Terahertz spectroscopy. Molecules with heavy atoms, such as some inorganic and organometallic substances, may absorb FIR waves, which induce to intramolecular vibration. Intermolecular stretching and bending of molecules with lighter atoms, and molecules with weak bonds such as Van der Waals contribute to differences in solid material inner structure and christalinity (polymorphism), which is successfully analyzed by FIR (Chalmers and Dent, 2006). The Mid infrared region (MIR) was the most popular and widely used infrared region in organic chemistry for long time. At first, its primary area of use was qualitative analysis, such as determination and detection of organic constituents and functional groups of unknown mixtures (Reeves and Zapf, 1998). Recent papers have reported successful quantitative applications in food, textiles, pharmaceutical and agricultural fields, but the use of MIR for quantitative use is not as developed and mature as for Near Infrared (Richardson and Reeves, 2005). The overall evolution of quantitative MIR may have been limited by (1) the relatively lower spectral

reproducibility when compared to NIR, and (2) the exigencies on thin samples due to the extremely high absorptibities of organic materials in that region (Wilson and Tapp, 1999; Smith, 2002; Chung et al., 1999, Rogo et al., 2007; Prieto et al., 2009).

The near infrared region (NIR), discovered by Herschel in 1800, is the most energetic infrared region and it is close to the visible region in the electromagnetic spectrum. His experiments in measuring the heat produced by filtering the sun light on colors with a thermometer, lead him to realize that temperature increased from going to blue (450 – 475 nm) to red (620 – 750 nm). Temperature kept rising even after positioning the thermometer further from the visible red, which meant that more energy was present beyond that which was visible (Herschel, 1800). Further significant research on the NIR region was not done for 150 years. MIR, meanwhile, became popular in analytical chemistry, while the NIR region was ignored as it was considered to lack relevant chemical information. The NIR spectra (absorption measurements in function of wavelengths) showed broad and overlapped low intensity bands, between 10 and 100 times more attenuated than sharper MIR fundamental absorptions (Dryden, 2003). NIR broad peaks could not be directly assigned to specific chemical compounds or interpreted in a straight-forward manner as MIR spectra. Later advances in computation and statistical methods helped overcoming those difficulties. NIR spectroscopy (NIRS) is currently more mature in quantitative analyses than MIR technologies and has at least two outstanding advantages: (1) NIRS allows measuring by reflectance (reflected light from a sample), so thicker samples with minimum preparation can be analyzed. (2) NIRS can pass through glass and optical glass fibers (Choquette et al., 2002) so measurements far from the spectrometer are possible.

1.2 Theory behind NIR Absorption

Light radiation has both wave and mass properties (wave-particle duality) because it is made of small particles or energy packages called photons. Some compounds can absorb light at certain wavelength or frequency, leading to changes in its atom energies. In order for a molecule to absorb a photon from the IR region, its molecular vibrational frequency must match the frequency of the IR radiation. Furthermore, its dipole momentum must

change during the radiation: The radiation and the molecule must interact in the way that the dipole of the molecule changes in the same direction as the electric field vector created as result of the radiation**.** Molecules that do not have dipolar moment, for instance homonuclear compounds oxygen or nitrogen, do not absorb IR light. The result of this light-molecule interaction is stretching vibrations that affect bond length in two-atom molecules, bending vibrations that affect bond angle in molecules with three or more atoms, and molecular rotations (Davies, 2005). The energy required to be absorbed varies on the bond length and the kind of vibration; for instance, according to Davies (2005), stretching requires more energy than bending.

The quantum theory states that molecules and atoms can only be found in states of certain energy. For being in a new state, the atom/molecule needs to absorb or emit energy equal to the difference between the first state and the new one. Because light energy depends on its wavelength and frequency, the emitted/absorbed light will have varying wavelength and frequency depending on the energy involved in achieving the new state.

Figure 2 shows the states associated to the electrons from an atom or molecular bond. Electron motion is the main responsible of the energy states thus the energy states are also called electronic states. The ground electronic state has the lowest energy and it is called equilibrium state in which the probability to find an atom or molecule is the highest in regular conditions. According to figure 2 and quantum theory, high energy ultraviolet waves can induce the electrons to “jump” to higher energy levels (called second electronic excited states), whereas absorptions in the infrared region (IR) induce changes in the vibrational states within the ground state (lowest energetic states).

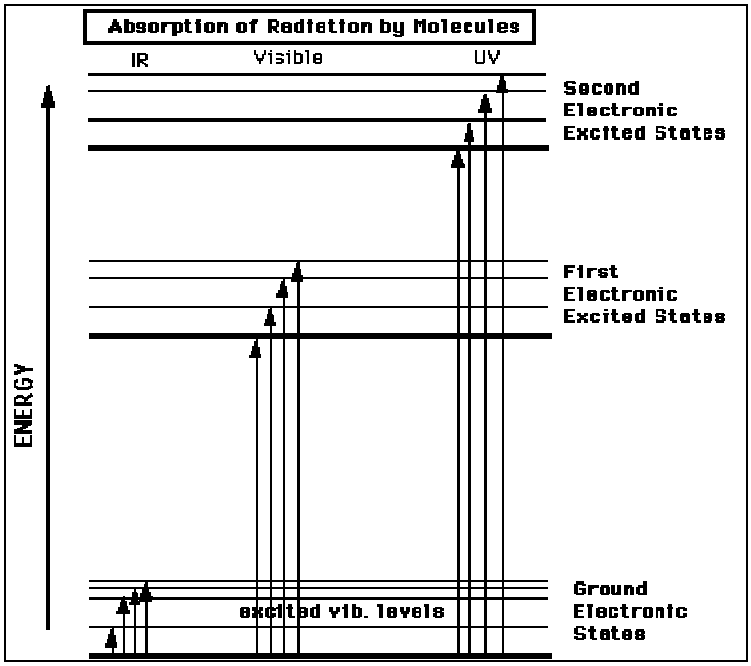


Figure 2. Energy levels for absorption of radiation (source: University of California, Department of chemistry)

One of the popular approaches to calculate the frequency of vibration (or energy of a state) of a diatomic molecule is using the harmonic oscillator theory or Hook's law (Eq1). There are two constraints when applying this principle: First, the quantum mechanics theory (Eq2) states that only discrete energy amounts can be absorbed - also called "allowed vibrational energy levels" -. And second, the *Selection Rule* which limits the transitions or "electron jumps" only between two consecutive energy levels. The absorptions that lead to these allowed transitions are known as fundamental absorptions.

Equation 1. $$\nu = \frac{1}{2 \cdot \pi \cdot c} \sqrt{\frac{k(m_1 + m_2)}{m_1 \cdot m_2}}$$

Equation 2. $$E = (n + \frac{1}{2}) \cdot h \cdot \nu$$

k= Force constant
C = Speed of light
n = Principal quantum number
υ = frequency of vibration
m_1, m_2 = nucleous mass from atom 1 and 2 respectively
E = Energy
h = Plank's constant

The principal quantum number in equation 1 is related to the energy level or electronic state and the shape of the orbital that contains the electron. When n= 1, the molecule is in its lowest energy state (ground state), the equilibrium state when there is the highest probability to find the molecule in regular conditions. If light excites an electron from a molecular bound such as n=2, we call it a fundamental absorption. As n increases, we find higher states of excitation and more energetic waves are required to achieve them.

The left plot on figure 3 shows the resulting parabola of plotting the potential energy function from the harmonic diatomic molecule versus the distance between atoms. It can be noted that according to the selection rule, the distance between energy levels is constant. The Franck-Condon principle of anharmonicity and Morse's potential function introduced relevant concepts to understand the existence of NIR absorption. Those principles account for Coulombic repulsion forces between atomic nucleuses and kinetic properties of atomic absorption as sources of hanarmonicity. The direct consequence of these phenomena are energy increments higher than the previously stated for lower energy levels in equation 2, and more stable levels (smaller energy increments) at atomic distances closer to the bound break-up point. Summarizing, the energy increments are not constant in reality and the updated potential energy function with Franck-Condon principle and Morse's function slightly differs from the harmonic approach as shown to the right on figure 3.

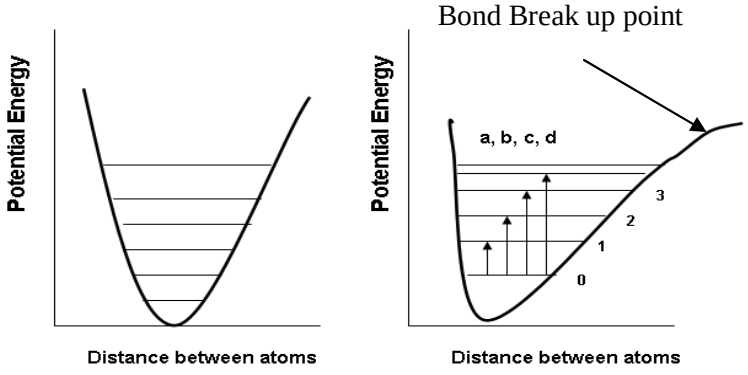


Figure 3. Harmonic Potential energy function (left). Potential energy function after Franck-Condon principles: a) fundamental absorption, b) first overtone, c) second overtone, d) third overtone

The explanation of overtones is another consequence of the anharmonicity conditions. Overtones are the result of bound absorption from ground state to higher non-consecutive energy levels: transitions higher than one energy state. This phenomenon was not initially contemplated by the selection rule, which stated it as a “forbidden transition” and it has small probability to happen in reality (a small fraction of the molecular bounds of a compound will experience overtone absorptions) and that translates in more attenuated signals when recording absorptions. The entire NIR spectra contains up to four overtones (although the fourth overtone is too weak to be measured) from the absorptions of the following groups: methyl C-H, aromatic C-H, methylene C-H, methoxy C-H, carbonyl associated C-H, N-H from primary and secondary amides, N-H from amides (primary, secondary, and tertiary), N-H of amine salts, O-H (alcohols and water), S-H and C=O groups (Workman, 2005). Note that all those bounds and groups are found in molecules that are part of organic matter and water. Absorptions involving hydrogen dominate in the NIR spectra because it is a light atom, thus it easily achieve higher vibrational transitions (Davies, 2005). Combination bands can be found at the highest NIR wavelengths, from 1900 to 2500 nm, and basically involve the same chemical groups as the overtones. They are the result of interactions between molecular vibrational frequencies, overlapped information from Fermi resonances, and inactive MIR bounds among other phenomena (Bokobza, 1998)

1.3 Using Radiation in Spectroscopy

When a sample is irradiated with light, according to the energy conservation law part of it is reflected, other is transmitted and another fraction is absorbed. The proportion of each depends on both the light wavelength/frequency and sample properties. The amount of absorbed energy is related to the sample composition (the compounds which absorb at that given wavelength and their concentration in the sample) as well as its thickness, as the absorption law (also known as Beer’s law) states:

Equation 3. Absorbance = $C_x \cdot e \cdot l$ C_x = Concentration of the test solute
e = Molar absorptivity of the test solute
l = Path length traveled by the light

Some of the assumptions that underlay the equation make its use very restricted, and barely impossible of being directly used in NIR spectroscopy. Some of the most restrictive assumptions are the consideration of no correlation between absorbers in complex matrices, homogeneous media: negligible light scattering, low analyte concentration in sample, and constant pathlength (Ingle and Crouch, 1988).

The light reflected fraction is favored when the sample compounds have high absorbance coefficients at the incident light (low compound absorbance) and the sample surface offers mirror-like characteristics. Reflection has dependence on the smoothness of the material's surface relative to the wavelength of the radiation. A rough surface affects both the direction and coherence of the reflected wave, hence it determines both the amount of radiation that is reflected back and the purity of the information that is preserved in the reflected wave. Light irradiation which is perpendicular to a smooth sample surface produces specular reflectance: the light is reflected back keeping the same perpendicular angle. When the sample surface has relevant impurities related to the incident wavelength, a high proportion of light is scattered in all directions and/or reflected as diffuse reflectance. There are two types of light scattering. Rayleigh scattering is non-directional and happens when the size of the particles is on the order of the radiation's wavelength. Mie scattering is due to larger particles and it is not as wavelength-dependent as Rayleigh scattering (Ingle and Crouch, 1988).

Reflected and transmitted light are the two alternative working modes in NIRS because as previously stated, direct absorbance measurement by Beer-Lambert law is not possible with NIR radiation. Variability in path lengths, non-homogeneity of samples, and scatter phenomena are few of the most common factors that make the law not directly applicable. Transmittance (T) is the ratio of the power of radiation after crossing a sample over the initial radiation power (P/P_o), expressed as % of transmitted light. It follows a theoretical straight forward relationship with absorbance (Equation 4), which interpretation is simple:

Sample compounds that may absorb strongly at certain wavelengths will not transmit so much energy through the sample, and the opposite is also true.

Equation 4. *Apparent Absorbance* = $\log(P_o/P) = \log(100/T(\%))$

Transmittance would be relatively easy to measure if scattered and reflected light were negligible, which is not always the case. Log(1/T) is also known as optical density, and although is not exactly the same as absorbance in usual situations due to the reflectance and scattered light, it still follows linear relationship with the real absorbance and it is successfully used in NIR spectroscopy.

Diffuse reflectance can also be related to the sample absorption and thus lead to a quantification of the analyte concentration of a sample using NIRS, while specular reflectance does not provide much information about the sample and spectrometers try to minimize it with the right alignment of detectors and filters. When the incident NIR beam is projected to the sample, part of it is scattered and transmitted several times by the sample particles so that the light collected by instrument sensors try to capture the re-emerging back-scattered radiation which has some degree of attenuation after sample absorption. Only the part of the beam that is scattered within a sample (proportion that reaches a sample depth dependant on the wavelength of radiation according to Hruschka (1987) and returned to the surface is considered to be diffuse reflection useful for analysis in NIRS. The best spectral region to work in diffuse reflectance mode ranges from 1200 nm to 2500nm, because below 1200 nm wavelengths are very energetic and absorption is weaker. There is high absorption above 2500 nm.

The most common way that diffuse reflectance is relating to absorbance with the following relationship (equation 5).

Equation 5. $$Abs = \log_{10}\left(\frac{1}{R}\right) = -\log_{10}(R)$$

In order to use the equation, a reflectance standard is required (a blank, such Teflon or spectralon with approximate reflectance equal to 100%), and the equation is applied as shown in equation 6.

Equation 6. $\log(R_{standard} / R_{sample}) = \log(1/R_{sample}) + \log(R_{standard})$

$\text{Log}(R_{standard})$ is constant, so apparent absorbance is reduced to one single term, $\text{Log}(1/R_{relative})$. With this transformation, the lowest value of absorbance a sample will reach will be 0 and the highest 2.

In many cases, an algebraic function called the Kubelka-Munk function is also used due to the qualitative resemblance of the output to the absorbance spectrum when the sample is thick, opaque and its absorption is weak (Cortat, 2003). The equation expresses the reflected light in function of two variables: dispersion (scattering) and absorption. The equation resulting from this theory is also ideal and it shows variability due to the difficulty to discriminate between absorption from the analyzed compound and the sample as a whole (matrix absorption) (Pou, 2002).

Equation 7. $\frac{k}{s} = \frac{(1-R)^2}{2R} = \frac{A}{s}$

R= Reflectance
k= Absorption coefficient
s= Scattering coefficient
c= Concentration of the absorbing species
A = Absorbance

K and S absorption and scatter coefficients have the inconvenient that do not have any physical meaning involved, thus the physical characteristics of a sample are not taken in account.

While sample pathlength is predetermined and must be kept constant for transmittance measurements, the minimum sample required in reflectance mode is highly dependent on the wavelength range used in the analysis and sample characteristics such as density or packing, particle size, and material absorption (Bertnsson et al., 1998). Physical characteristics affect reflectance measurements especially at higher wavelengths

(combination bands region) hence any sample changes will create an additional source of variability and noise in the measurements (Norris and Williams, 1984).

Overall, reflectance measurements show shorter dynamic range compared to transmittance (lower sensitivity) because information provided by diffuse reflectance originates from smaller sample portions and has been attenuated (Corti et al., 1999). Its repeatability is slightly worse which is more noticeable in heterogeneous samples. In specific applications, those limitations may not create significant errors, or may be mitigated by using wider range of wavelengths (Kays et al, 2005). Transmittance measurements exceed the accuracy of reflectance measurements in most pharmaceutical applications, although analytical sensitivity, signal to noise ratio and limit of detection is highly affected by sample position and changes in geometry (Short et al., 2008). Comparison studies in agriculture fields do not lead to a unanimous consensus regarding superior performance of any of the two measurement modes (Williams and Sovering, 1993; Borjesson et al., 2007). Although there is a general preference towards transmittance measurements when small concentrations need to be measured, differences arise from in combination of factors such as selected wavelength range, instrument and sample characteristics, data processing/analysis, and sampling procedure (Kays et al., 2005; Short et al., 2005; Delwiche, 1995; Cogdill et al., 2007).

The signal collected from either reflectance or transmittance after irradiating a sample is graphically displayed by most of the instruments as a plot of wavelength (usually in nanometers) on the coordinates axis versus the absorbance value on the abscise axis. The measurements are connected by a line, giving as a result a spectrum plot. Figure 4 shows an example of NIR spectrum and the approximate location of the overtone and combination bands regions.

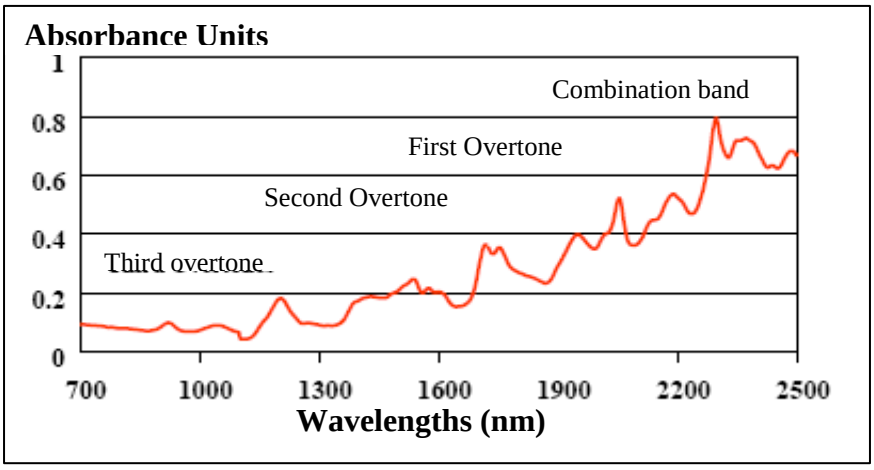


Figure 4. Example of NIR spectrum with the overtones and combination bands zone

2. Instrumentation: Spectrometers

Despite proprietary instrument conformations, any commercial NIR spectrophotometer has five basic sections further detailed: (1) Sample compartment, (2) Light source, (3) Light wave selection system, (4) detector/s, and (5) signal processor or computer.

2.1. Sample Compartment

Instruments working by reflectance do not need sample confinement for in-line measurements, but it is common to use open sample cups or sample cells confined by silica or quartz (materials transparent to NIR light) in laboratory instrumentation (Figure 5). Transmission instruments may work with confined sample cells as well, but with specific pre-set pathlenghts ranging from 0.1 to 10 cm, depending on the product to be analyzed. An integrated adjustable sample compartment with automatic flushing is used for whole grain analyzers. One of the advantages of NIR light is its ability to pass through optical glass fibers preserving most of the signal integrity (losses lower than 5% per km of cable), even if the resulting output intensity is low. This is especially useful for measurements to be made far from the physical instrument and for multiple sampling/ sequential analyses in multiplexer systems. The use of optic fibers with probes for either transmission or diffuse reflectance measurements allows sampling by immersion in

liquids for controlling fermentation or other liquid reaction processes (Buchanan et al., 1988; Tamburini et al., 2003; Sarraguca et al., 2009), contact on small sample areas such as works of art (Bacci et al., 2005), in-vivo medical analysis (Yu et al., 2007), and development of smaller spectrophotometers (Smith, 2000).

Figure 5. Sample cells and compartments of NIR of NIRS instruments

2.2. Light Source

The most popular NIR light source is the tungsten halogen lamp, which has wavelength emission ranges from 320 to 2500nm. The halogen gas allows recycling of the evaporated tungsten (Stark and Luchter, 2005), and brings the advantage of longer lifetime compared to traditional tungsten lamps without halogen.

Light emitting diodes (LED) were used as light source in the first commercial instrument for whole seed analysis in 1985 and in the first portable spectrometers (McClure et al., 2002). The low power consumption, price, small size, and long lifetime (around 25 years) of LEDs still make them the most suitable light sources for miniaturized instruments and specific screening applications outside the laboratory environment (McClure et al., 2002, Axun technologies, 2005). Conventional LEDs emit in short wavelength ranges (30 – 50 nm) around their center point. Several of them can be mounted in an array with narrowband interference filters if wider wavelength ranges need to be covered, although

measuring many wavelengths with this configuration is not an economical approach (Malinen et al., 1998). LED devices have been improved during the recent years to overcome some of their limitations. For instance, some commercial instruments allow easy switching of LEDs according to the application.

Finally, the most innovative light sources are tunable diode lasers, or also called superluminiscent light-emiting diodes (SLED). Using the semiconductor technology of diodes, tunable diode lasers are much smaller than the traditional tunable laser, cheaper, with excellent wavelength resolution, brighter, and with lower noise frequencies than tungsten lamps. SLEDs are suitable for measuring weak absorptions at good signal-to-noise ratio and as light sources in miniature instruments. Improvement of tunable diode lasers allow controlling emitted light at specific wavelength, combining light source and wavelength selection features.

2.3. Wavelength Selection

Most detectors collect light intensity from a relatively wide range of wavelengths. Recording signal values at specific wavelengths is required for analytical purposes. Filters were the first wavelength selection device to be used for this purpose and are the element in the spectrometer which leads to more diversity of instruments in the spectrometer market for infrared spectroscopy (Stark and Luchter, 2005). The most simple filters work by absorption (absorption filters), which are discrete bandpass filters that absorb all light wavelengths but the one of interest.

Narrow bandpass interference filters (Fabry-Perot) achieve better spectral resolution and higher output intensity by selecting wavelengths according the refractive index and thickness of the dielectric material between the two layers of reflective material (Pou, 2002). To select multiple wavelengths, interference filters are mounted in a wheel which can be automatically controlled to rotate and select the suitable filter for the wavelength selected. This creates spectrometers that provide few spectral measurements and are usually called filter photometers instead of spectrometers. Although filters are an alternative that provides acceptable results, problems of image misalignment and slow operation are common. (Balas, 2009).

Acousto-optic tunable filters (AOTF) and liquid crystal tunable filters (LCTF) allow faster tuning for wavelength selection, and provide better reproducibility without the need of mechanical devices because one filter can generate several output wavelengths. AOTF filters modulate the light wavelength and intensity through the interaction of sound waves generated in a birefringent TeO_2 crystal. The frequency of the acoustic signal makes the refractive properties of the crystal change allowing wavelength specific transmission. Wavelength discrimination in liquid crystal tunable filters (LCTF) is carried out by applying variable voltage to progressively change the polarity of a liquid crystal (Garini et al., 2006). Those filters provide a better output quality compared to AOTF filters, but their short wavelength range is limited (below 1800 nm), and give a lower intensity dependent on the selected wavelength (Stark and Luchter, 2005; Balas, 2009).

Dispersive type instruments use a prism or a grating, which diffracts the incident collimated light beam at different degrees while resolving it in discrete wavelengths. Light dispersion can be done before scanning a sample (predispersive instruments) or after radiating the sample with polychromatic light (postdispersive). Postdispersive instruments offer advantages such as less environmental interferences with the lamp radiation, analyzing wider sample areas, and hold longer distances between sample and light sources (Schumann and Meyer, 2000; Wang and Paliwail, 2006). Prisms have been replaced by gratings because of lower cost and better linear wavelength dispersion of the last ones. There are two types: Holographic (photosensitive film with fringes) and ruled (concave surface with fringes). Ruled gratings require being complemented with other optical elements such as lens, and show less stray-light rejection than holographic gratings (Holler et al., 1998; Domanchin and Gilchrist, 2001).

In the dispersive instruments group, there are monochromators and spectrographs such as diode-array instruments. Monochromators are pre-dispersive instruments that scan a sample with grating mechanical motion. The basic principle is as follows (Figure 6): Polychromatic NIR light enters through an entrance slit and is then collimated (light rays are made parallel) by a mirror. The light hits the dispersion grating and later hits a focusing mirror, which reflects it to a second exit slit to either hit the sample (transmittance mode) or hit the single-channel detector (reflectance mode). Entrance and

exit slits of a monochromator are very carefully designed to have accurate geometry since they are critical for instrument observed resolution (smallest wavelength difference distinguished by the spectrometer) and effective wavelength bandwidth (full width of a band at half of its maximum value, FWHM). When using grating alone without slits, resulting resolution is not enough for most chemical measurements in plastic or pharmaceutical applications (Thermo Fisher Scientific, 2006). Small slits (around 0.1 mm) give low band width, more dispersion, and high spectral definition useful in qualitative applications; large slits (around 2 mm) give more intense radiation and are more suitable for quantitative analysis (Holler et al., 1998).

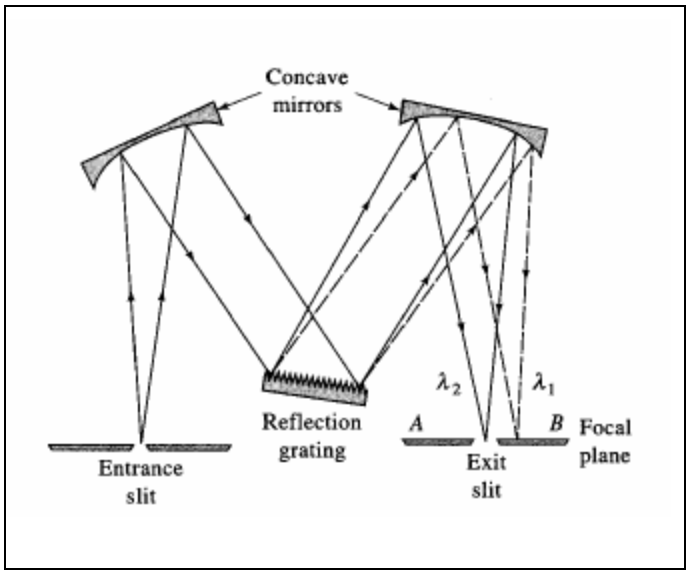


Figure 6. Schematic of a monochromator predispersive instrument

Diode arrays spectrographs are post-dispersive instruments that measure all the wavelengths at the same time thanks to a fixed grating and a set of detectors placed in array (multichannel detectors). There is no need of exit slits. There are fewer optical elements compared to monochromators and resolution depends on the number of elements in the detector array and array characteristics. The latest advances in wavelength selection besides tunable light sources are the Micro-Electro-Mechanical Systems (MEMS) created with semiconductor technologies. MEMS diffraction gratings control light diffraction by electronically controlled movement of diffracting

microelements. Their small size and lower cost has lead to a new generation of portable instruments. Figure 7 shows one of the pioneer NIR hand-held instruments in using MEMS as grating technologies, Phazir by Polychromix™.

Figure 7. Handheld spectrometer (Phazir) by Polychromix™

2.4. Detectors

Detectors transform the incident light energy to electric analog signal. The electrical signal is then amplified and transformed to digital, which may later be further processed by the computer. Detectors and amplifiers are considered the most common sources of non-systematic noise in instruments (random noise). Random noise is reduced in most commercial instrumentation by averaging several spectra from a same sample, improving the signal-to-noise ratio (SNR). SNR achievable values in NIR spectroscopy according to Workman range from 25,000:1 to 100,000:1 (Workman, 2005).

An effective detector must have a linear relationship between the energy input and signal output within its dynamic or working range - from the minimum detectable signal to the maximum before reaching saturation -. Measurement linearity is influenced by other

factors besides detector characteristics; for instance, the number of bits of the analog to digital converter device and slight detector misalignments, which can lead to capturing a small fraction of the reflected specular component (often called stray light) in reflectance mode instruments. Without linearity, more complex and potentially unstable mathematics are needed to calibrate the instrument.

Photo-sensitive detector materials are chosen according to the NIR region to be covered. From 400 to 1100 nm, silicon detectors (Si) are common (Stark and Luchter, 2005). Si detectors are stable, fast, not too expensive, and sensitive to low light intensity to achieve good performance. Lead Sulfide (PbS) or Indium Gallium arsenide (InGaAs) detectors can cover higher wavelength regions than Si detectors, being usual working with both types. Photodiode Arrays (PDAs) spectrographs have a set of InGaAs detectors or charged coupled devices (CCDs) in array (Figure 8). While InGaAs PDAs offer high signal precision, require less signal processing, are simpler to build, have high SNR and less sensitivity to high light intensities when compared to CCD, CCDs have higher signal sensitivity and resolution (Greensill and Walsh, 2000). CCDs are usually used for imaging devices later explained. PDA instruments take faster measurements (all wavelengths measured at the same time) and can be smaller in size than grating monochromators, which optical conformation cannot be easily reduced in size because it would lead to low throughputs and resolution (Smith, 2000).

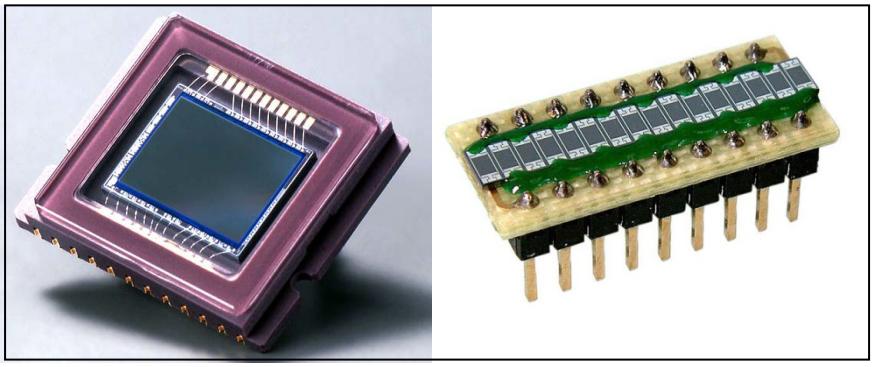


Figure 8. Pictures of a charged-coupled device (CCD) (left) and a photodiode array (right)

3. Other Near Infrared Technologies

There are other NIRS technologies and instrumentation use of NIR light under slightly different principles from traditional spectroscopy. Two of the most well-established are Fourier transform NIR (FT-NIR) and NIR chemical imaging, further discussed.

3.1. Fourier Transform NIR (FT-NIR)

Fourier Transform (FT) is widely popular in MIR spectroscopy, and it has recently gained high popularity in NIRS. FT technology offers advantages such as high SNR, high light outputs due to absence of slits, fast measurements, instrumental simplicity, and high resolution and accuracy (Thermo Fisher Scientific, 2006). Brimmer et al. (2001) claim that those advantages are more perceptible when working in the MIR region due to the limitation of higher detector noise relative to signal when working in the NIR region.

FT-NIR measurements are carried out in time domain and the direct instrument output from sample scanning is an interferogram instead of a spectrum. NIR interferometers (Figure 9) split the NIR light beam in two; one of the beams is reflected to a fixed mirror, and the other is reflected to a mirror that moves forward and backward at carefully controlled speed - usually tuned by a HeNe laser-. The reflected beams are recombined back in the beam splitter to generate the interferogram signal, which is a result of light interferences. When displacing the moving mirror, the pathlength difference respect to the fixed mirror change, leading to different grades of interference between the two reflected beams and which are correlated with different light frequencies. After the interferogram light reaches the sample, transmitted or reflected signal is read by the detector in time sequence (ms), hence measurements are fast. Although interferograms contain information from all the frequencies or wavelengths encoded, it has to be first processed with the Fourier transform. The computation takes as an input a time domain wave signal (the interferogram) from which the transform principle states signal is made from an addition of sines and cosines of a set of individual wave frequencies. The processed signal or output looks like the spectra obtained by any traditional spectrometer, but with the expectation of higher throughput and frequency accuracy. One of the drawbacks is the fact that FT-NIR instruments are complex and expensive, and suitable

for controlled environments mainly (such as laboratories), due to their sensitivity to external factors such as temperature and vibrations.

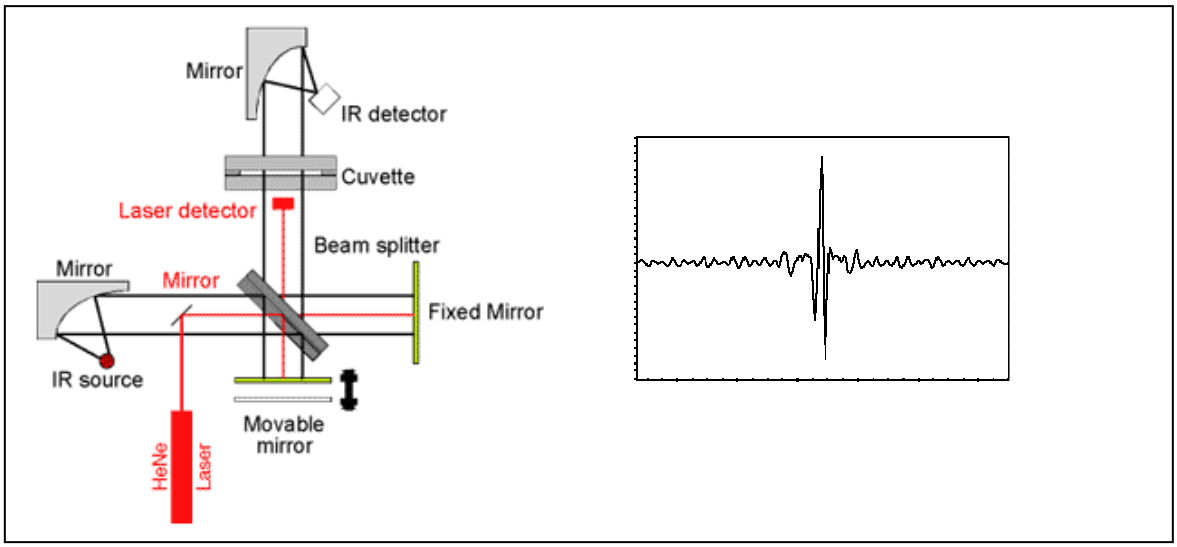


Figure 9. Scheme of a Fourier-Transform set up of optics and mirrors (left). The result is an interferorgram (right). Image retrieved from Foss instruments website.

3.2. Hyperspectral Imaging

Near Infrared chemical imaging (NIR-CI), or also called NIR hyperspectral imaging, has rapidly become popular, especially in measurements by diffuse reflectance. It combines the advantages of near infrared spectroscopy with digital mapping: the chemical compounds of a sample can be both discriminated and quantified in the sample spatial frame. This is especially useful for analyzing compound distribution and sample heterogeneity. Instrument parts and operating principle are very similar to traditional spectrophotometers. The sample scanning procedure can be carried out in two ways: 1) by push-broom or moving imager technique, popular for in-line measurements and sensing, or 2) by fixed staring systems.

Pushbroom instruments measure a spectrum from a whole sample by small consecutive areas or lines while the sample platform is moved and their wavelength selection is usually by dispersion. Staring systems scan on still samples, one wavelength at a time,

using either AOTF or CLTF filters. The mapping capability of imaging systems is brought by digital cameras with 2 dimensional arrays of detectors (pixels) such as CCDs that are effective in lower light intensities. Pixel size or area analyzed per pixel range 49 to 1,600 squared microns in commercial instruments, depending on selected magnification. Higher magnification (or smaller sample area captured per pixel) will lead to more detailed spatial analysis and a lower dilution effect of the compound of interest within the sample matrix.

NIR-CI Data Structure

NIR-CI data structure can be thought of as a cube or a stack of cards, where two spatial dimensions are combined with a third dimension corresponding to the chemical information or spectra (wavelengths). Depending on the manufacturer, around 320 x 512 pixels (2D) are arranged to capture both sample area and spectra. In that previous example, a total of 320 x 512 = 163,840 spectra would be generated for a single wavelength and correlated to small sample portions as a chemical map. If the instrument had 200 sampling wavelengths, the final “image” or data cube would have a total of 320 x 512 x 200 = 32,768,000 data points. Although the amount of data generated is large, visual selection of image areas or pattern recognition techniques help discarding pixels with no relevant information.

This concept is illustrated in figure 10, where each squared surface is like a picture taken at one single wavelength and the small squares within represent pixels. In common imaging terminology, “samples” and “lines” specify the number of columns and rows of pixels; “bands” refer to the discrete number of wavelengths, or following the previous analogy, the number of cards in the stack.

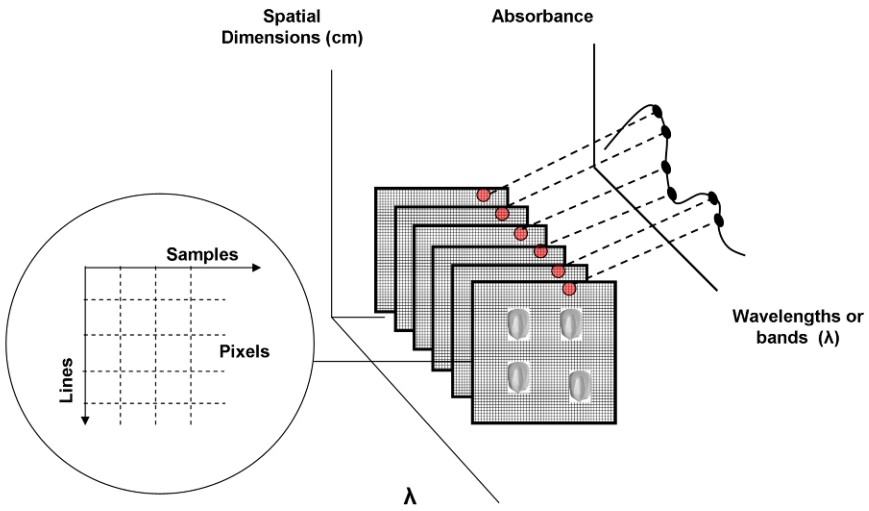


Figure 10. NIR-CI data configuration, showing the data cube of images at different wavelengths. A single spectrum is obtained from each pixel.

Captured data or pixel readings are usually recorded as analog to digital sensor units (ADU), also known as raw units. Modern instruments digitize with more than 16 bytes of resolution. Digital data can be stored following different interleave types or formats such as band-sequential (BSQ), band interleaved pixel (BIP), or band-interleaved line (BIL). The main difference among these formats is the order and sequence that pixel readings are stored: BIP for instance stores first in a single file the information from a same pixel at different wavelengths or bands (store the whole spectrum obtained from that pixel), while in BSQ formats the whole image at a certain wavelength is first stored, and then the second, as a deck of cards so all images are stored as separate files. Since in NIR chemical image applications there may be over a hundred wavelengths to be stored, BIL and BIP formats are preferred to avoid too many data files.

Data are stored in an image general format. ENVI is a common raw data format for hyperspectral images which has two files: One file contains the binary data, and another file is a plain text header that provides information about the data such as the interleave storage format, data dimension, or sensor specifications among others. It is important to

specify the type of data that has been stored in a file. Data can be stored as raw counter units or normalized/reflectance units. Raw counter units are directly readings from the camera sensor that must be digitized and are not useful for spectral applications. Raw data are function of the light intensity and, at higher degree, a function of sensor sensitivity which is dependant on the wavelength, as it is the detector dark current (Geladi et al., 2004). Pixel calibration or data normalization is needed to correct for dark noise and a posterior transformation to reflectance units.

Similarly to the normalization procedure carried out on single point instrument data, a reference standard of high reflectance is used to calibrate the data at single pixel level. But before pixel calibration, readings taken without sample are intended to provide dark current data that is removed from the sample images. Most of the instruments, for time and expenses reasons, are calibrated using a single Teflon reflectance standard of 90 - 99% reflectance. Knowing the reflectance value of the standard and obtaining the raw counter values, each pixel can be calibrated to reflectance units using the following one-point linear regression (for a standard of 100% reflectance value and for per cent reflectance units):

Equation 8.

$$x_{reflectance} = \frac{(Sample - Black)}{(White - Black)} \cdot 100$$

The best pixel calibrations are carried out when several standards of different reflectance are available because they allow non-linear calibrations (such as quadratic), which is found to successfully model the normal detector reading behavior (Geladi et al., 2004).

4. Chemometrics

NIRS data analysis requires chemometrics. Chemometrics, a term widely used in NIRS-related literature, refers to the use of mathematics, statistics and computational devices in chemical analysis. Without computing capabilities and multivariate methods, NIRS applications would not be possible. Chemometrics made possible the dealing of NIR

units in resolving highly overlapped and broad peaks, high sensitivity to sample physical characteristics, and high information redundancy. While information redundancy can be an advantage and can allow working with different wavelength ranges, determining which wavelengths hold information of interest without having correlation between them is not a problem which can be efficiently solved by trial-error experimentation.

Figure 11 shows a block diagram with the basic steps for developing a NIRS calibration. In that procedure, the broad absorptions (spectra) from a sample irradiated with NIR light are correlated with the compound concentration or sample characteristic which user pretend to analyze. The compound to be measured should either be of organic nature (direct measurement) or be correlated with sample physical characteristic or another organic compound (indirect measurement). Some relevant aspects of the calibration procedure can be pointed from the diagram: 1) there is the need for a fundamental analytical method, called the reference method, in order to obtain the dependent variable to be calibrated; 2) a suitable number of samples uniformly covering a wide enough range of analyte concentration should be part of the calibration set, and 3) the calibration model should be later validated to test the model performance on future samples.

Chemometric methods are at least used in two stages: preprocessing of spectra and model development. Outlier detection is another stage which may require chemometrics depending on the data complexity. Sophisticated multivariate methods such as genetic algorithm or particle swarm optimization may be also used for selecting the variables to be included in the calibration or discrimination models. Some of the most common preprocessing methods, calibration and discrimination methods are further explained.

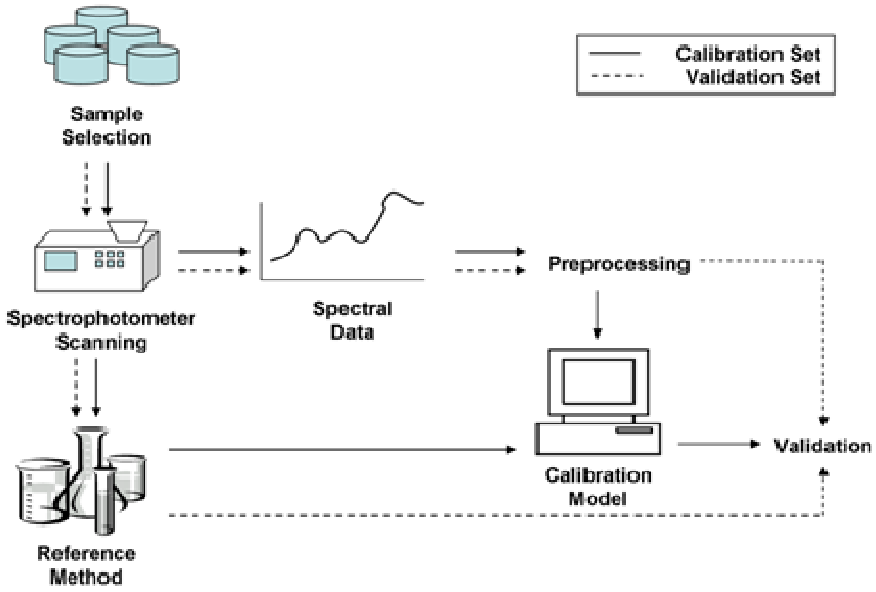


Figure 11. Diagram of steps for developing a NIRS calibration

4.1. Selecting Samples, Reference Methods, and Spectral Data

The importance of choosing an adequate calibration set is often underestimated and not usually covered in the literature. There is no fixed number or rule-of-thumb to determine the number of samples to be included in a calibration. At least between 20 and 30 samples should be taken for feasibility studies and initial calibrations (Williams, 2001), but more robust calibrations may use few hundred (for instance, instrument built-in calibrations for grain analysis). Calibrations of homogeneous mixtures (i.e. pharmaceutical powders) may require smaller calibration sets than agriculture samples of high compositional complexity and heterogeneity, such as whole grains or forages.

Users work under the constraints of sample availability and reduced budget. Nevertheless, there is not enough emphasis on the ultimate consequence of using calibrations developed with inadequate calibration sets: calibrations with low predictive ability. An ideal calibration set should cover the chemical, spectral, and physical characteristics of the population to be analyzed and avoid future extrapolations when predicting new samples (Fearn, 2005). The distribution of reference values should be uniform. If the distribution

is normal (bell shaped distribution), samples belonging to either higher or lower concentrations have the chance to get more relevance in the calibration, which would not be desirable. Because reference values are not always known and reference analyses of large sample sets may be expensive there are other methods to select an initial calibration set, using spectra. A method developed by Naes (Naes, 1987) and later illustrated by Naes et al. (2002a) uses principal component analysis (PCA) on the spectra and cluster analysis of the data. PCA is a technique that projects the spectral data to a new reduced dimensional space and it is later explained in detail.

NIRS calibrations can match or virtually achieve better precision and accuracy than traditional wet chemistry methods (Coats, 2002), but paradoxically, NIRS relies on them for calibrations. The quality of the reference data influences NIRS calibrations. Careful search for the suitable method and laboratory should be carried out. In the case of NIRS instrumentation for grain analysis, calibrations are often preloaded, e.g. wheat protein. Although this may seem an opportunity to save time and resources in developing custom calibrations, the performance of any built-in calibration must be carefully validated to determine its suitability for a particular situation. Calibrations from an instrument brand and model may not perform successfully when loaded to a similar instrument, or used on different samples than the original calibration population. Outliers from either reference values or spectral data exist and most calibration methods are highly sensitive to them (Kovalenko et al., 2006; Hubert et al., 2008).Visual check of the spectra can identify abnormal and noisy spectra. Visual check is often not enough, and possible outliers may not be detected until data is either preprocessed, or a first attempt of calibration has been carried out.

Detecting multiple potential outliers is not simple; their effect is masked with each other. Traditional approaches to detect single outliers do not perform well (Walczak and Massart, 1998, Naes et al., 2002b). The use of influence measures such as Leverage or Hotelling's T^2 statistic in combination with checking model residuals are a powerful alternative for outlier detection (Haaland and Thomas, 1988). Influence statistics give an idea of how different a sample is from the rest of the data in a given dimension. While that does not explicitly make a data point an outlier, high leverage followed with a high

residual value (the sample was poorly modeled by the model) give chances that the sample is an influential outlier. Its exclusion from the calibration set could improve the calibration. However, if removed, enough similar samples should remain in the calibration set to avoid significant reduction of representativeness, especially in reduced data sets.

4.2. Spectra Pretreatments

Pretreatments or spectral preprocessing methods are a set of mathematical procedures on spectra before developing a calibration model. Although raw or not preprocessed absorbance values can be used directly to create good predictive models, other times the nature of the samples require mathematical preprocessing for better model performance. Mathematical pretreatment of spectra reduces noise or background information (smoothing techniques) and increases signal from the chemical information (differentiation). In other words, they allow overcoming variability in sample thickness and differences in light scattering, keeping a more linear relationship between analyte concentration and absorbance values. Any applied pretreatment must lead a robust model with good predictive ability. Basically, preprocessing methods can be classified as baseline correction – normalization, signal enhancement, and statistical filtering of signal noise.

The selection of best pretreatments depends on the signal and data origin, such as instrument and sample characteristics, but the selection of the best method usually requires trial-error and user experience. Although there are several techniques explained in the literature, most of them are variants from the basic well-known pretreatment methods following explained. More than one method can be used simultaneously in any order, although any scatter correction technique (MSC, SNV, normalization) should be performed prior to differentiation techniques. Users should avoid increasing the complexity of models using too much preprocessing, when in fact, the opposite should happen (i.e. fewer latent factors in Partial Least Squares models).

4.2.1. Mean Centering

Mean centering is carried out subtracting the average from all spectral values at each individual data point from each spectrum, moving the mass of the data center to the space coordinates origin without affecting the distance between the points. When performing this operation, we remove the absolute absorbance or baseline value so the data analysis focuses more on the absorbance variability in each wavelength. This basic pretreatment is commonly used when we carry out principal component analysis (PCA) or partial least square regression (PLS) later explained in detail. Centering the data to the mean value can reduce the model complexity, often reducing the number of latent variables to be employed by one (Haaland and Thomas, 1988).

4.2.2. Scaling

Scaling gives as a result variance equal to 1 for each wavelength after dividing each individual value by the standard deviation of all the values in that wavelength. Scaling the whole data matrix allows each wavelength to have the same weight (or eliminates the initial weights) during the modeling procedure, which is suitable if the previous relevance of variables for the calibration is unknown. Haaland and Thomas (1988) suggest not using scaling when the errors are independent from the changes occurring in the spectra (error not being proportional to wavelength) and most of the spectra do not contain much chemical information, since the importance of variables containing noise will get the same importance as variables containing chemical information. When operations of scaling and mean centering are performed together, the operation is called autoscaling (or Z transform), and while the data spread is affected by any type of scaling, the relative data distribution and overall meaning is not affected (Lavine, 2000).

4.2.3. Mean Averaging Smoothing

Mean averaging, or best known when applied as the moving window mean averaging, is a filtering method that performs smoothing after calculating the average from the data

points inside the window (which length is user-specified) and replacing the value of the first data point with the mean. The process is done moving the window one data point and carrying out the same procedure until the end of the spectra. Other modifications to this smoothing technique is the weighted mean moving window, which tends to give more weight to the central point in the data window, or the median smoother. Using the median smoother instead of the mean helps to remove data spikes but is not as effective removing noise (Alfasi et al., 2005).

4.2.4) Multiplicative Scatter Correction (MSC)

The method, applied by Geladi et al. in 1985, helps to eliminate sample physical characteristics – path length variability- from the spectra. Packing and geometry differences in reflectance spectra can lead to offset baseline due to increase of light intensity from specular reflectance (Gemperline, 2006).

It is proven that the representation of a sample spectrum versus the average spectrum from a set of samples (which is used as a basis or reference spectrum) give almost straight lines (Figure 12). When applying MSC, the spectra is first averaged and each individual spectrum is regressed by partial least squares to the total average. The regression equation slope and intercept represent the additive and multiplicative effects of light scattering, respectively. Finally, each spectrum is corrected for offset (the offset value is subtracted) and each wavelength of the spectrum is divided over the slope. The regression coefficients should be stored and applied to new data. The variability of points from the line is interpreted as the variability due to the chemical information. Each individual spectrum is corrected removing the offset value from the whole spectrum and dividing it by the slope.

Figure 12 shows three spectra from three different samples represented versus the average of the set they have been selected from as a black line. The sample with pink spectrum and pointed with an arrow shows the most additive scattering effect.

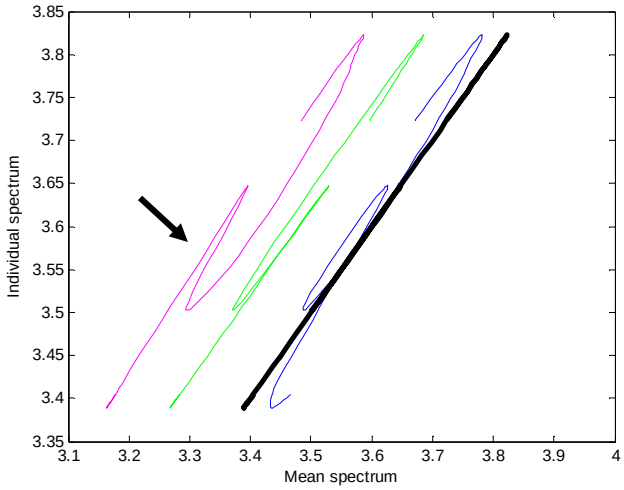


Figure 12. Sample spectra vs average spectra of all the samples

MSC is advised to be applied in wavelength regions where there is not much chemical information, which is often difficult to determine in NIRS, or on whole spectrum when the scatter effect is important. Otherwise, the technique will remove also information related to the chemical composition together with the scattering effects (Naes et al., 2002c). Pretty recent publications suggest a loopy application of MSC. Performing MSC more than once allowed correction for additional variations in several sets under experimentation (Windig et al., 2008).

4.2.5. Standard Normal Variate (SNV) and Normalization

The method, developed by Barnes et al. in 1985, centers and scales the spectrum of each sample. The mean of the spectrum is subtracted from each spectrum wavelength (in equation 9, *a* is equal to the spectrum average value), and the result is divided by the standard deviation of the spectrum (b in equation 9) so the total set of spectra has a mean of 0 and variance equal to one after the treatment. SNV removes the offset and any effect that contributes to the overall variation of the spectrum. It is common to use de-trending techniques which account for the baseline shift and curvilinear. The results after this transformation are similar to MSC, but with the advantage of not needing a reference

spectra and not requiring to save regression coefficients. The application of both techniques requires a careful checking of score plots for data artifacts and non-linearities that may arise after the data transformation.

Equation 9. $$\lambda_{corrected} = \frac{\lambda_i - a}{b}$$

Normalization techniques are similar to SNV, but *a* from equation 9 equals to 0, and b is a vector-norm. For instance, the most used is the Euclidian Norm (the square root of sum of the squared elements).

4.2.6. Derivatives

Derivatives enhance relevant peaks correcting for overlapping as is shown in figure 13. The models resulting from applying derivatives usually require fewer factors, thus resulting models are more robust. Although it is possible to use high degree derivatives, up to fourth degree derivatives are the most used in the literature; first and second derivatives are the most common. First derivative removes the displacements from the baseline (constant factors for all the wavelengths) while second derivative corrects the terms that vary linearly with the wavelength (Pou, 2002). Forth degree derivatives are found to be good to curve sharpening and good for absorbers separation (Hopkins, 2008). Gap-segment and Savitzky-Golay are the best known techniques to perform spectra derivatives pretreatment. It previously requires fitting a polynomial function to the spectra. Fitting the spectrum to a certain polynomial function can be used as a smoothing technique without applying derivatives as well. Savitzky and Golay polynomial involves the creation of data subsets from a spectrum, and fitting a polynomial by least squares. The process is done for each data subset of user defined number of odd points (or also called data window). The window is moved one point to the right to fit again the polynomial of user-defined degree. Higher degrees would allow a better fit, but then no smoothing action would be performed and the applied derivatives on the polynomial

would enhance the noise. Choosing the window size for the polynomial fit is also important, since big windows can remove both noise and signal, and small windows may not improve the signal.

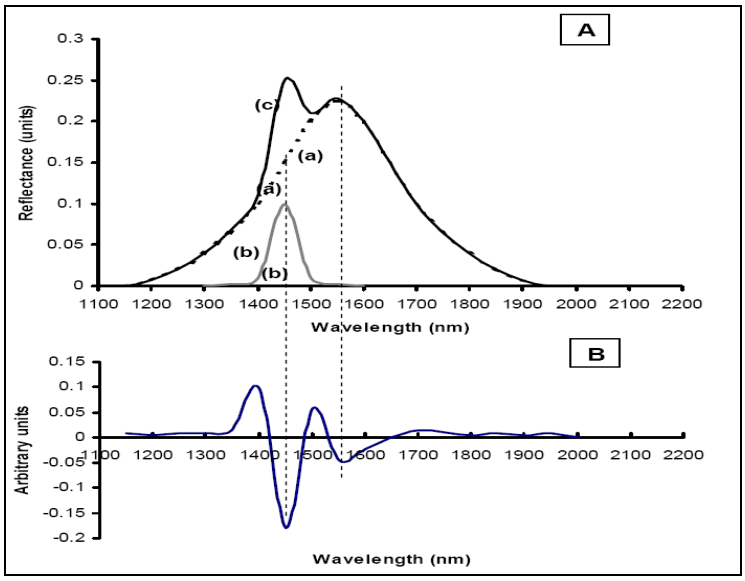


Figure 13. Spectra with overlapped peaks (A) is corrected with second derivative treatment (B), where the individual peaks arise. (*Source: Drydden, 2003*)

4.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is independently here because its popularity and involvement in most of the NIRS analysis. Besides being used in combination or coupled with other regression or discriminatory analysis, the same principle with some modifications is applied in methods such as PLS.

PCA is well known in clustering analysis and data compression. Spectral data is usually formed by a high number of variables (wavelengths) with a high correlation degree. That is to say, the information provided by several wavelengths may be redundant and show collinearity. When regression is carried out on highly correlated variables, fitted regression points do not provide a defined structure and the resulting model is unstable.

For instance, if regression is carried out with two variables highly correlated (Figure 14), the points fit a plane (doted lines) which spans low variability.

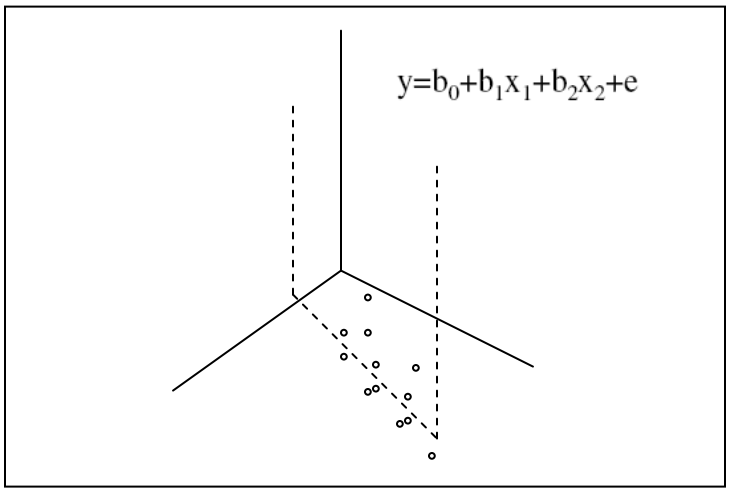


Figure 14. Plane formed when carrying out a multiple regression with two highly correlated variables

Basically, PCA summarizes the variance-covariance matrix of the spectral variables, reducing the dimensionality of the data but keeping the main information from the variables. Geometrically, PCA changes the initial highly correlated axis of the data to a smaller set of axis, called principal components (PCs). The data is projected on new orthonormal axes (PCs which are perpendicular with each other and unit length) which are built as linear combinations of the original variables: the wavelengths. The algorithm finds these new axes seeking for the orthogonal directions which explain the maximum data variability. The first PC will be drawn following the direction which explains the highest variability. The second PC will seek the second direction of maximum variability under the constraint of being orthogonal (perpendicular) to the first PC. The third PC will seek the third direction of maximum variability being perpendicular to the first and second PCs and so on. The PCA concept is represented in Figure 15. The initial data (a) is plotted in 3 dimensions where two of them are highly correlated. Step b shows the way that PCs would be created: PC_1 follows the direction of highest variability as a combination of the two highly correlated variables. PC_2 is generated through the next

direction with more variability and at the same time the angle with PC_1 is 90^0. PC_3 would mainly represent residual variability, and could be omitted in (c), where data has been rotated and fitted to the new axes or PCs, eliminating data collinearity.

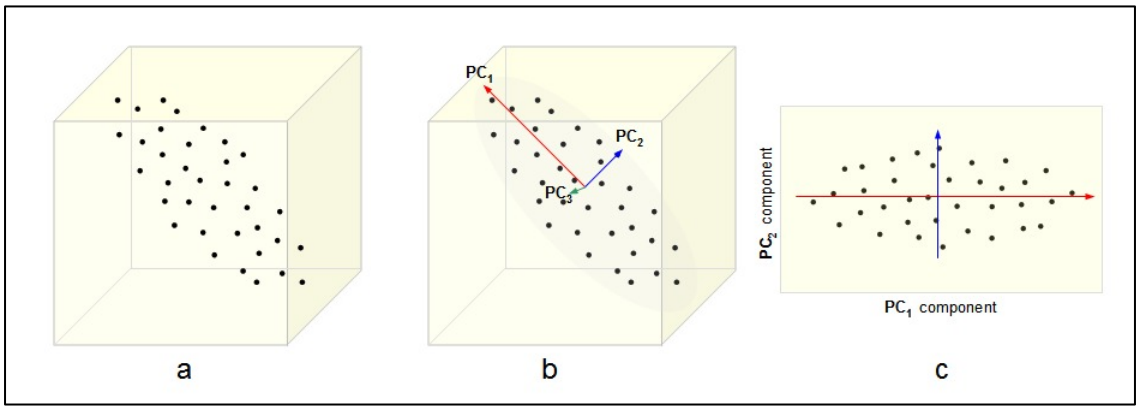


Figure 15. Example of Principal Component Analysis (PCA) on three dimensional correlated data (*Source: Kavraki, 2007)*

The new axes or PCs are defined by the loadings, which are the cosines of the angles that each PC forms with the old axis (wavelengths). The loadings can also be seen as the weights for each original wavelength in each principal component.

Previously to carry out PCA, data must be normalized by autoscaling or mean centering so the data is centered. The original data will be projected to the new PC axes according to equation 10.

Equation 10. $T = X*P + E$

X is the original data matrix. P is the matrix of loadings (P). T is the score matrix, or the new values that original data acquire on the new axes. E is the matrix of residuals. The number of PCs that can be calculated depends on either the number of initial variables or samples, but commonly only up to 20 are calculated from NIR data. Only few PCs – the first ones - are considered important at the end, depending on the variance they explain

over the total data variance. This can be checked from the eigenvalues. Each PC has an eigenvalue associated (the sum of all the eigenvalues is equal to the number of PCs) which is a constant number obtained through the projection and calculation process of each PC (eigenanalysis). The last PCs will not have important information because they explain very small sources of variability, usually associated with noise. When researchers use PCA to summarize their data, they may use all PCA that have an eigenvalue higher than 1, or commonly, they plot all eigenvalue and do not take further PCs after the first inflexion point of the plot or elbow.

4.4. Linear Calibration Models

Sample absorbance is expected to be linearly related to the compound to be measured according to Lambert's law in most of the cases. Three of the most popular linear calibration methods are following discussed. Among them, Partial Least Squares (PLS) seems to be the most popular although its performance is similar to principal component regression (PCR).

4.4.1. Multiple Linear Regression (MLR)

MLR is one of the oldest multivariate regression methods, used by Norris in his later experiments. The method provides good results if the number of measured wavelengths is relatively low (for instance, data measured from filter instruments) or an advanced wavelength selection method such as Genetic Algorithm could be applied. Another important consideration when using this method is to avoid strong collinearity among the wavelengths: the information from each of the wavelength measurements is not strongly correlated with any of the others.

MLR is a generalization of the univariate inverse method based on least squares fitting of *y* to ***x*** **(**it is also known as inverse least-squares). The algorithm gives a linear regression equation (equation 11).

Equation 11. $$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + error$$

Each independent variable (for $i = 1 \ldots n$) x_i is correlated with the dependent variable (the reference value) and its correlation is measured with the coefficient of correlation r (or coefficient of determination r^2). This is done in a stepwise manner through creation of a sequence of multiple linear regression equations unless it is known beforehand which variables are going to be introduced. At each step of the sequence, one variable that makes the greatest reduction in the error sum of squares of the sample data (or the one that provides the greatest increase in the F statistic) is added to the regression equation. The process is continued until some stopping criterion is met or all the predictors are processed. In this manner all possible linear regressions on all subsets of the available independent variables are tested. The subset of predictors that produces the lowest standard error is reported. The error term is also known as residuals. One of the problems associated with MLR is that it is prone to over-fitting (Davies and Grant, 1987), when a significant amount of irrelevant information (noise) or too many predictors are incorporated into the model.

4.4.2. Partial Least Squares (PLS) and Principal Component Regression (PCR)

Both PCR and PLS successfully deal with wavelength correlation and redundancy of information. PLS, as it is following explained, can be seen as an improvement of PCR since the principal components are calculated not only taking in account the spectral data matrix but also the reference values. For this reason while PCR is considered an unsupervised regression method, PLS is already classified as supervised regression.

PCR is a direct application of the principal component analysis (PCA) method. Once the spectral data is projected to the new orthogonal non-correlated dimensional axis (PCs) a regression process by multiple linear regression least squares is performed between the projected data and the reference values. Wold's introduction of PLS (1975) was an improved alternative to PCR; Both methods carry out regression on data projected to a new dimensional space, but the new space coordinates created in PLS regression take in account the information from the reference value matrix. For this reason the new variables receive the name of latent variables (LVs) instead of principal components

(PCs). The way that the latent variables are calculates is by maximizing the covariance between X and Y. That is to say, X and Y are decomposed at the same time and while the covariance between them is forced to be maximized; another constraint sets that both matrices residuals from Y and X decomposition are close to zero. Both data and reference matrices are mean-centered or autoscaled previously of being decomposed as follows in equation 12 and 13, respectively:

Equation 12. $$X = TP^T + E$$

Equation 13. $$Y = UQ^T + E$$

Where T is the matrix of scores, P is the loading matrix for the spectral matrix. U is the score matrix and Q is the loading matrix for the concentration matrix. The P loadings are not following the exact direction of maximum variability since they are also considering the information from the reference matrix: The decomposition of both matrices is not independent; it is done simultaneously so there is an inner relationship described in equation 14:

Equation 14. $$\hat{u}_a = b_a t_a$$

Where for each component *a* a regression coefficient b establishes the relationship between the scores of the spectral and reference matrices. Through this relationship, the reference matrix can be expressed as in equation 15:

Equation 15. $$Y = TBQ^T + Residuals\ in\ Y$$

Where B is the matrix with the regression coefficients. The prediction equation can be expressed in a way that new data does not need to be projected again in the latent variables. Being x_i the spectrum from a sample of unknown concentration, this can be predicted using the following equation 16:

Equation 16. $$y_i^T = b_0^T + x_i^T B$$

Where b_0 is the intercept and B is the matrix of the regression coefficients.

Although both methods provide similar results, PLS become more popular. PLS accuracies may not usually be significantly higher than those of PCR but they are achieved by including fewer latent variables in the final calibration (Naes et al., 1986; Hammateenejad et al., 2007; Muñiz et al., 2009). PLS is preferred because the algorithm is faster, models have higher precision, and provides more harmonious calibration models (Kalivas and Gemperline, 2006). There are at least two main algorithms to perform PLS calibrations: NIPALS (Non-linear Iterative Partial Least Squares) and SIMPLS. NIPALS works slower but is told to be more transparent than SIMPLS, which is faster (Wise, no date).

PCR and PLS calibrations are only based on a relatively small number of PCs/LVs because similarly to what has been explained with PCA, since they are extracted following the direction of maximum data variability, the last PCs/LVs usually involve noise. If an excessive number of variables are included in the calibration, a fraction of noise is also modeled and the calibration becomes too specific to the calibration set. This phenomenon is known as overfitting and leads to a reduction of model accuracy in future predictions. There are different approaches to estimate the appropriate number of PCs/LVs to be kept for the calibration model. One of the most employed approaches uses cross-validation, mentioned later as an alternative validation method as well. The general idea of cross-validation is to keep a single sample (full-cross validation or leave-one-out cross-validation) or a group of samples (k-fold cross validation) apart and develop a calibration with the remaining samples. The remaining samples are then predicted by the developed calibration (validation) and the prediction values are compared with the real reference values to calculate the error. This procedure is consecutively done until all the samples have been predicted once. The error is finally expressed as Predicted Residual Error Sum of Squares (PRESS) (equation 17, where h_{ii} are the leverages for each

observation). In other words, PRESS is the addition of the squared error from each sample when predicted by the model.

Equation 17. $$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(\frac{residual}{(1 - h_{ii})} \right)^2$$

The PRESS value can be used to select the number of latent variables or principal components in the final model. Chemometric software does this cross-validation procedure using several PCs or LVs and displays the cumulative PRESS value graphically so users may visually select the number of PCs/LVs that lead to the first minimum PRESS value from the plot. The best number of PCs to select would be the one that shows the elbow on the plot as then the value does not significantly decrease.

Another option is calculating the ratio of the new PRESS after adding a new PC/LV over the residual sum of squares before adding the new latent component since it gives an idea of the significance of the new component added (i.e. how much variance is explaining), considering it significant when the ratio is equal or smaller than 0.9 (Wold et al., 2004). The F test by Haalan and Thomas (1988), is based on the idea of picking the component number which PRESS value is not significantly higher than the previous one. The ratio of PRESS value at certain component over the PRESS at the previous component is calculated for all the components, and it is assumed to follow an F-Fisher distribution with degrees of freedom equal to the number of observations. Each ratio is compared with the tabulated value for the one-tail F distribution to determine when two PRESS values are not significantly different.

4.5. Non-linear Models

There may be cases where the relationship between sample spectra and reference values is not linear. There may exists several sources that make the relationship chemical information-spectra nonlinear, such as stray light, detector's characteristics, very high or low absorbance, overlapping signals or particle size (Despagne and Massart, 1998). Any

of the previously cited calibration methods can handle small nonlinearities, but when prediction residuals show certain pattern of positive and negative values or plots of predicted versus reference values show appreciable curvature, nonlinearity need to be addressed (Naes et al., 2002d). Often, curvature may not be noticeable but for the fact that calibration statistics are not good. The Durvin-Watson statistic can be used as an assessment tool for detecting nonlinearity (Howard and Workman, 2005). Although being a low power statistic and requiring a large number of samples to be used, it is a sensitive and specific test for nonlinearities in predictions. The test is based on the ratio of standard deviation by successive differences of residuals by the ordinary standard deviation of residuals. Its value is close to 2 (not significantly different than two) when the residuals are not correlated – no hidden non-modeled variability source or nonlinearities -. It assumes that ordinary standard deviation is dependent on the curvature, so if nonlinearities exist the denominator will increase and DW will achieve values significantly lower than two. To perform the test, data must be first ordered according the reference values.

Once the nonlinearity is detected, there are some solutions suggested by Naes et al. (2002d) such as new preprocessing, deleting wavelengths, adding extra principal components/latent variables to the model, using nonlinear calibration models, or split the data in subsets using an approach similar to the cluster analysis for sample selection (Naes, 1991). If none of those approaches work, there are nonlinear methods that handle more complex relationships between spectra and analyte. Those are following explained.

4.5.1. Local Regression and Locally Weighted Regression

Local regression (LR) and locally weighted regression (LWR) are not intrinsically nonlinear regression methods –local methods use subsets of the whole data-, but since they allow developing traditional calibration methods such as PLS or PCR in non homogeneous or highly clustered data they allow modeling nonlinearities up to certain extent.

LR and LWR are applied piece-wise. This make the method useful for highly clustered data since approximately one local regression will be developed for each cluster independently. Since the local models are built with smaller sets of samples (neighbours) there is some risk of having not very stable calibration parameters (Despagne and Massart, 1998) and in order to use this method with accurate results is recommendable to keep adding samples to the calibration pool.

One of the LWR algorithms is found in the PLS_Toolbox, a set of functions developed by Eigenvector Inc. for use with Matlab™. A global calibration model is performed first on the calibration set (i.e. PCR, PLS) and the loadings are obtained. During the prediction of an unknown sample, its score on the first PC is calculated. The distance between the sample and the rest of the data pool in that PC is calculated (Mashalanobis distance), finding the closest set of used-defined number of neighbors. For the closest neighbors, the distance of all the points in the neighborhood is normalized dividing by the radius of the neighborhood (i.e. the largest distance value), and a weight function is applied to give more relevance to the closest points for the prediction of a new sample. There are several available weighting functions, but the tricube (equation 18) seems to be the most popular and used in the algorithms. Linear and biquadratic can be obtained substituting the cubic exponent in equation 18. Engster and Parlitz (2006) mention that results obtained from the different weighting functions in time series analysis is pretty similar in terms of prediction accuracy, but the higher the exponential term the fewer points affect or contribute to the local model.

Equation 18. $$W_S = \left(1 - d^3\right)^3$$

where W_S is the weight associated with a calibration sample and d, the scaled distance between the new sample to predict and the calibration sample.

Summarizing, LWR develop a model with training data, but when new predictions must be done, the training data is retrieved again. It could be defined as a “dynamic” model. Although LWR takes in account the whole set of data every time a sample must be predicted, it only uses for prediction the set of data that is close to the sample and that is

the reason why it is called “local”. It does not need a high number of initial training samples when compared to Artificial neural networks (ANN) explained later, although it requires more samples than PCR and PLS (Burns and Ciurzack, 2007).

4.5.2. Artificial Neural Networks Regression (ANN)

Artificial Neural Networks (ANN) is a computational method that can be applied to NIR data to develop nonlinear calibrations. By trying to simulate the human nervous system, ANN uses the calibration set to learn about any relationship (no matter how complex and does not need to be linear like in MLR, PCR or PLS) that may exist between spectra and references. ANN regression is much more complex that the previously mentioned methods and require adjusting and optimizing several parameters. The most common type of nets for this purpose, which we will later describe a little more, are the multi-Layer feed-forward backpropagation learning nets.

An artificial neural net is composed by neurons (the basic units) or nodes, layers, and transfer functions that join the neurons from different layers (Figure 16). When working with NIR spectra, the input nodes would be either the absorbance values from the wavelengths or the scores from principal components, and the output node would be the predicted value. Other nodes may be created in hidden layers (multilayer perceptron model), which increase model complexity and ability to model non-linear relationships.

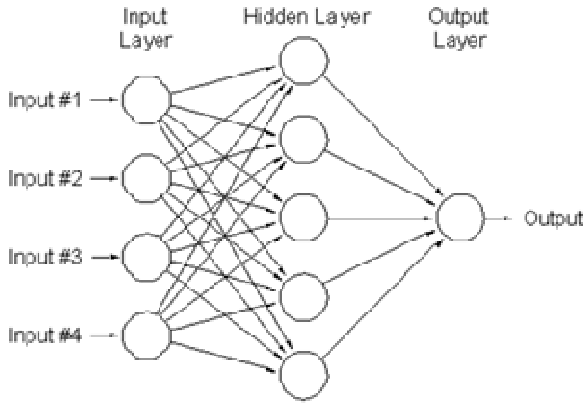


Figure 16. Structure of a neural net with 4 inputs, 5 hidden neurons in a single hidden layer, and a single output

The nodes or neurons, similarly to biological neurons, have a specific number of connections with other neurons by the transfer functions. The connections acquire specific weights during the training process, which either positive or negative values (excite or inhibit). The absolute value of the weights will depend on the relationship between the input data and the target or output. The weights will define the efficiency of information transfer to that connection, and can be interpreted as a way of stored knowledge. The neuron input is the result of the dot product of the input vector and the weights vector, which is then processed by the activation function (equation 19).

Equation 19

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

Where j is the input connection, w is the weight of the connection j for a neuron k, and x is the input value coming from the connection j.

Therefore, the output of the neuron would be the outcome of the activation function on the value of v_k. Among the several activation functions cited in the literature (threshold, piece-wise linear, and sigmoids) the sigmoid type are the functions used for multilayer models, with common functions such as the logistic and hyperbolic tangent (Massart et al., 1998). Sigmoid functions have linear response for intermediate values, and can model nonlinear responses thanks to their non-linear behavior in their extremes (Figure 17). The hyperbolic tangent function forces the results to be in the range of -1 and 1 (normalization), but both curved tails offer a lack of sensitivity during training of highly non-linear sets and require more time to modify the weights (the tails are activated to model non-linear responses). The linear functions are most common in the output layer if most of the modeling part has been carried out in the hidden layers and are preferred for the output layer in classification problems (Massart et al, 1998).

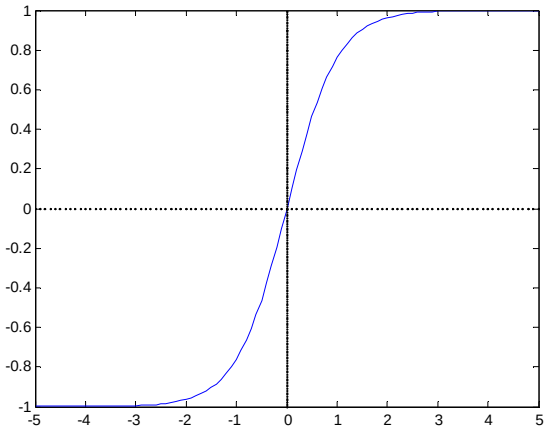


Figure 17. Graphic representation of hyperbolic tangent function

The nodes or neurons are located in layers. Nets with hidden layers with more nodes interconnected performing functions are known as multilayer perceptron model, although direct connection between input and output layers is possible (no hidden layer). The number of nodes or neurons per layer should never exceed the number of training samples, and usually half the number of training samples is proved to work well (Despagne and Massart, 1998). The optimal network morphology (number of hidden layers and neurons) can be determined growing the network (adding additional neurons and carrying out training) while observing the network performance. Once there is no more improvement, no more neurons are added. On a similar way, we can start with a big number of neurons and removing one until no improvement is observed; this process is called pruning. Usually with just one or two hidden layers the network performs with accuracy. The output layer can have more than one neuron or node, that is to say, multiple responses can be modeled although unless the responses are correlated to model one output at a time is proved to perform better (Despagne and Massart, 1998).

Previously to ANN, PCA is usually applied for data compression if the number of variables is large, although other methods of data compression such as wavelets and Fourier together with a wide variety of variable selection methods can be used to keep the variables with high relevance for a given prediction or classification problem.

Learning Procedure and Delta Rule

Once the net morphology has been set, it needs to be trained. During the training is when the weights between connections are set and gradually modified under either supervised training (when targets are provided) or unsupervised training (the same input spectra is provided as target to make the net classify spectra by their patterns).

There are several ways to carry out training. Incremental training performs modification of the weights every time an input is processed, so it is a slower training process but equivalent to the faster batch training method, which performs weight modifications after all the inputs are processed- it is thus a more memory consuming process -. These modifications are done following any of the existing learning rules. One of the most popular is the delta rule, which states that a weight is modified according to the difference between the obtained value from the activation function and the desired one (equation 20): Whenever there is a difference, there is learning.

Equation 20. $$\Delta w_{ki} = \eta\,(t_k - x_k)\,x_i$$

Where x_i is the input associated with the initial weight w_i, x_k is the current output from the activation function of the neuron k, *t* is the target, and η is a learning rate which value is between 0.0 and 1.0, as later explained.

The delta rule is based on the gradient descent theory. It assumes that the error function follows a n-dimensional parabola-shaped surface when representing the squared error versus the weight value. The surface has a minimum that ideally would be reached when obtaining the ideal weights through carrying out the delta rule learning. The ideal situation happens when there is a single linear activation function in the network without hidden layers. However, almost all the ANN applications require a hidden layer with one or more neurons. In such cases, the error surface does not follow the ideal parabola and several local minima may exist, forcing the training process to stop when the best weights are not yet achieved.

In order to train more complex multilayer networks, more than the simple application of the delta rule for learning is required. The backpropagation learning algorithm offers a

solution to train more complex methods under the similar principles of the delta rule – and for this reason the algorithm is also called the general delta rule -. The algorithm works assigning random weights in the beginning, processing the inputs through the network (additions and activation functions), and comparing the final output with the targets. The error is calculated using the squared error for instance, and it is backpropagated through the network modifying the weights (Equation 21).

Equation 21. $$E = \frac{1}{2} \cdot \sum_k (t_k - x_k)^2$$

Where t is the target value and x is the actual output value of the neuron k. The weight modification is done by derivation of each output error function in each neuron respect to its inputs, and it is evaluated respect to the weights (Equation 22).

Equation 22. $$\partial E / \partial w_{ki} = \partial \left[\frac{1}{2} (t_k - x_k)^2\right] / \partial w_{ki} = -(t_k - x_k)\, g'(h_k)\, x_i$$

From the equation, g(h) is the activation function applied to the h input. Sigmoid activation functions are derivable and continuous, one of the conditions that the algorithm requires in order to carry out this learning process (Massart et al., 1998). The error function derivative respect to the weight provides the weight modification value (Equation 23).

Equation 23. $$\Delta w_{ji} = \eta (t_j - x_j)\, g'(h_j)\, x_i$$

η is again the training rate. The learning rate of a network measures the change rate of weights in each iteration or epoch. The desired learning rate is neither too fast nor too slow in order to achieve a convergence with the goals without taking too long in computing. It depends on the activation function, while sigmoidal activation functions perform well at learning rates above 0.5, linear activation functions work with considerably lower values (Massart et al., 1998). High learning rates may lead to system oscillation that could be avoided introducing a momentum term in the equation, which basically control which proportion of the calculated weight difference will be taken for

the modification. The number of iterations to be performed should not be excessive since the net would start modeling noise and becoming too specific for the training set, loosing accuracy for future predictions.

Network Performance

The best way to control the training process and network performance is through monitoring the training error after each iteration, together with the error of a monitoring set. The monitoring set is independent from the training set, thus the error obtained after the set is processed by the network is expected to be higher than the one obtained from the training set. Examining both training and monitoring error trends by plotting them help to determine when there is overtraining (overfitting or noise modeling that leads to poor model generalization) or the network is not successfully learning. In this last case, both errors do not decrease enough to be acceptable and we may want to check the transfer function values: This phenomenon tends to happen when too many nodes lead to have the activation function values either too high (approaching 1) or too low (Massart et al., 1998). Checking the error trend also allow finding system oscillations which may have learning rates too high.

The net training is stopped once the sum of squared errors from the monitoring set either stabilizes or falls below a given threshold, or the net performance degenerates. Because the number of initial samples to develop an ANN model is required to be high to get an appropriate training, the number of samples is usually a limitation for using this technique. An additional third sample set (test set) could be used to test the network predictive performance for new samples. Whenever it is possible, model validation should be referred to the test set; otherwise, the monitoring set could be used as validation set which would be an equivalent of the use of cross validation for PLS. The use of cross validation in ANN is also possible if there are not enough samples to set a monitoring set.

ANN is complex due to the big amount of parameters that need to be controlled and tuned. Furthermore, it has been considered a “black box” method. The interpretation of weights and judging the importance of the input neurons is difficult or nearly impossible.

4.5.3. Support Vector Machines (SVM)

The SVM method is based on principles of statistical learning theory developed by Vapnik and Lerner (Vapnik and Lerner, 1963). Initially, the method was intended for solving classification problems, but then was adapted for linear and nonlinear function estimation (Drucker et al., 1997).

SVM applied to a linear regression problem would be carried out using a regression function as the equation 24 (with < > denoting dot product), where x are vectors from the data matrix and w is a weight vector that defines the regression "hyperplane". The predictions should not exceed a specified deviation value respect to the targets ε (Figure 18), while the function should have small curvature through minimizing ½ $||w||^2$. This constraint helps to reduce the number of infinite functions that could be constructed through the given finite points (Durbha et al., 2007).

Equation 24. $F(x) = wX + b$

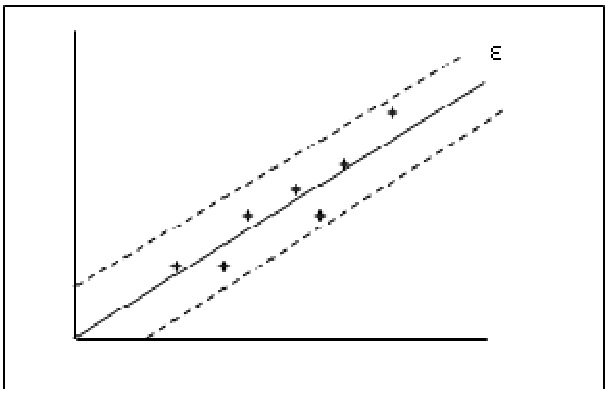


Figure 18. Regression by SVM and established maximum deviation between samples and regression line

Because there may not always be a function that can be built under all those constraints, the introduction of two new variables (C and ξ) in the function allows controlling the relevance of the constraints. How flat the function should be and the tolerance value of

exceeding ε is controlled with the regularization parameter C in equation 25 to be minimized. C allows controlling the global deviation of the model: small C brings more tolerance to the errors (more importance to obtain a flat function) while high-value C brings more complexity to the model while it becomes more sensitive to errors (Colliez et al., 2006) which could give more relevance to possible outliers.

Equation 25. Minimize $\frac{1}{2} \|k\|^2 + C \Sigma(\xi + \xi^*)$

ξ and ξ^* are slack variables, lower and upper difference from ε , which mathematically leads to the constraints (equation 26)

Equation 26. $y - w X - b \leq \varepsilon + \xi$ and $- y - wX + b \leq \varepsilon + \xi^*$
with ξ and $\xi^* \geq 0$

A loss function is defined to determine ξ and the quality of the model predictions. The linear ε–insensitive loss function (Vapnik, 1995) is the most commonly used since it does not account for ε but only for sample distances higher than ε. Other loss functions are quadratic, Huber, or Laplace. Furthermore, the linear loss function presents the highest robustness for estimations since it has the lowest rate of error increment compared to the other loss functions (Colliez et al., 2006). Equation 27 and figure 19 represent the linear loss function in mathematical and visual forms, respectively.

Equation 27. $L^{\varepsilon}(y) = \xi = \max(0, |y - f(x)| - \varepsilon)$

So $\xi = 0$ if $y\text{-}f(x,w) \leq \varepsilon$ and $|y\text{-}f(x,w)| - \varepsilon$ otherwise

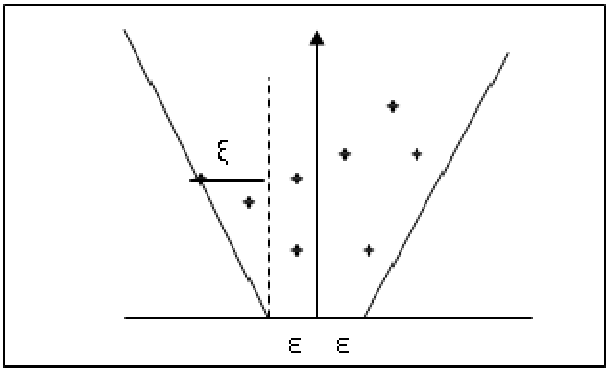


Figure 19. Graphic representation of the linear loss function. The perpendicular line locates the regression line.

It is important to notice that in order to work with the regularization parameter constraints, only the points which variability ξ exceeds the previously specified ε, as is shown in figure 19, contribute to the loss function. Those points are called support vectors and used for model development. If ε is set too high, it may lead not having support vectors to build the model, and on the opposite, with ε too small too many points are used in the model which can lead to a lack of the generalization ability. The slope of the loss function is specified by the regularization parameter C.

In order to find the optimal solution to the equation 27, Lagrange multipliers optimization can be performed to obtain the weight vector (w) (equation 28).

Equation 28. $$\mathrm{w} = \sum_{k=1}^{N} \alpha_k \cdot \boldsymbol{x}_k$$

This is one of the strengths of SVM compared to ANN: the Lagrange error function only has a single minimum thus there is no risk to be stuck in a local minimum (Zomer, 2004). The resulting function or regression model is shown in equation 29.

Equation 29. $$\hat{y} = \sum_{k=1}^{N} \alpha_k \cdot \boldsymbol{x} \cdot \boldsymbol{x}_k + b$$

where vector $\boldsymbol{x}$ represents new sample, $\boldsymbol{x_k}$ is k^{th} training sample, α_k is Lagrangian multiplier for k^{th} training sample , and b is the offset calculated from the Karush-Kuhn-Tucker conditions. There is an important observation from equation 29: the algorithm takes the dot product of the vectors of the data matrix, which means that the dimensionality of the data is "not important" in the function. This is the key that allows applying support vector machines for non-linear regression problems.

Until now, the model has been developed using linear correlation between input data and targets, but the same procedure can be used for non-linear correlation because the final optimum regression function is not based on data dimensionality. The same SVR algorithm can be used in highly dimensional data. When SVR is performed on the initial space, a linear correlation may not be possible but the regression may be linear in another highly dimensional space. To get the data to this new space, the initial data matrix is mapped using a mapping function Φ. Thanks to the fact that the regression function predicts using dot product of the data, there is no need to know about the mapping function details, but work with what is called a kernel function (equation 30).

Equation 30. $K(x_i, x_j)=\Phi(x_i)*\Phi(x_j).$

The kernel K function is equal the product of two samples in a given high dimensional space. Kernel functions do not deal with Φ implicity, so computations are less complicated and take less computation resources even if Φ is highly dimensional (Burges, 1998). Similarly to equation 29, we can introduce the kernel function in final regression quation (Equation 31).

Equation 31. $$\hat{y} = \sum_{k=1}^{N} \alpha_k K(\boldsymbol{x}, \boldsymbol{x}_k) + b$$

There are several well-known kernel functions such as polynomial, linear, sigmoid, splines, additive kernels, or inverse-multi quadratic. Linear kernel is the simplest one, the rest require tuning additional parameters related to the kernel characteristics. One of the

most used is the Gaussian Radial Basis function (RBF) (equation 32), because there is the need for only tuning one parameter (σ). It is of easier computation than polynomial kernel, for instance, that requires finding the best degree and an addition constant.

Equation 32.

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

The RBF kernel parameter (σ) is known as the kernel Pardzen window width, which affects the degree of generalization of future predictions. Small width will lead to have less variance and project to higher dimensions, but lead to higher bias as well. The best kernel parameters are chosen usually together with the optimal values of ε and C by grid search (by small increments of the kernel parameters and comparing the obtained error) and cross validation. The grid search allows simultaneous tuning of the parameters – since they can not be optimized independently – applying increments of each variable and comparing the prediction error from each combination and retaining the combination that minimizes that error. It is usually time consuming and not necessary to work over large ranges of values – furthermore, the risk of falling in local minima of the error function is high-, so other methods establish approximate values where the grid search could start from. Cherkassky and Ma (2002) suggest using the following approximation in equation 33 for obtaining the value of C.

Equation 33.

$$C = \max\left(\left|\bar{y} + 3\sigma_y\right|, \left|\bar{y} - 3\sigma_y\right|\right)$$

where ŷ and σ are the mean and standard deviation of the reference data, respectively. ε can be approached depending on the sample size (n) as shown in equation 34 (note that for big sample sizes another random constant τ needs to be defined by the user). Other alternative methods are based on pattern search (Momma and Bennet, 2002; Frohlich and Zell, 2005) but they have not replaced the popularity on the grid search-cross validation search.

Equation 34. $\varepsilon = \frac{\sigma}{\sqrt{n}} \quad if\ n \leq 30 \quad \Big| \quad \varepsilon = \tau\sigma\sqrt{\frac{\ln(n)}{n}} \quad if\ n \geq 30$

Least squares support vector machines (LS-SVM) is a variant of SVR that uses least squares instead of the ε–insensitive loss function. Instead of working with slack variables ξ and ε, the least squares function works taking the error from each data point to the regression function (Figure 20); The function to minimize this time is shown in equation 35.

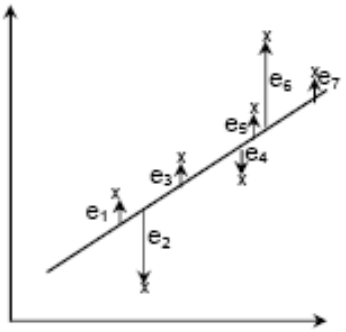


Figure 20. Sample error to the regression line

Equation 35. Min ½ $||k||^2$ + C/2 $\Sigma(e^2_i)$

With the equality constraint $y_i = x_i w + b + e_i$ for i=1 to n

Although LS-SVM uses all the training data points, thus all the points are support vectors, it is computationally faster than SVR because it avoids the quadratic programming needed to solve the constraints from the Lagrangian function in SVR. The problem is reduced to a linear system instead (Abe and Onishi, 2007).

4.6. Pattern Recognition and Classification

Pattern recognition is a widely used term to refer to the identification of clustering or similarities in data, with or without previous information about existing data classes. When there are established initial data classes, methods of supervised classification such as SVM classification or discriminant analysis can be used. Other methods such as K-neighbor allow data classification when classes are not known a priori (unsupervised classification).

Most of the previously explained calibration methods can be used for supervised classification. PLS in classification (also known as PLS discriminant analysis, PLS-DA) and ANN work under the same principles of regression models using dummy variables as reference. When regressing with dummy variables, the reference values are set in either a vector or a matrix with as many columns as many classes are, which has only one 1 per row (sample) in certain column, to indicate the class where each sample belongs. The predictions of new samples to be classified won't be necessarily a matrix of 0 and 1, thus the classification process from the predictions starts setting adequate thresholds. Thresholds are set according to the number of samples in each class, setting it as the division of the number of calibration samples from a certain class divided by the total number of samples. However, it is important to note that methods such as ANN can be biased if the number of samples in each class significantly differs.

SVM was initially designed for linear classification. As classifier, its objective is to draw a line or hyperplane that has a maximum distance to objects on the border of the classes, that is, maximizing the space (also called margin) between samples on the boundaries from the two classes and which are the support vectors. The optimization process (maximization of the margin) is carried out using Lagrangian multipliers similarly to the optimization of the slack variables in SVM regression. The classifier gets the final form of equation 36 where y_i gets values of either 1 or -1 depending on the class it belongs.

Equation 36. $$f(x) = sign(\sum_{i=1}^{i=n} y_i \cdot \alpha_i \cdot x_i \cdot x + b)$$

Note that the function relies again on dot product so kernel mapping functions allow classifying the data in highly dimensional spaces when the linear separation in the current dimensions are not possible (Figure 21).

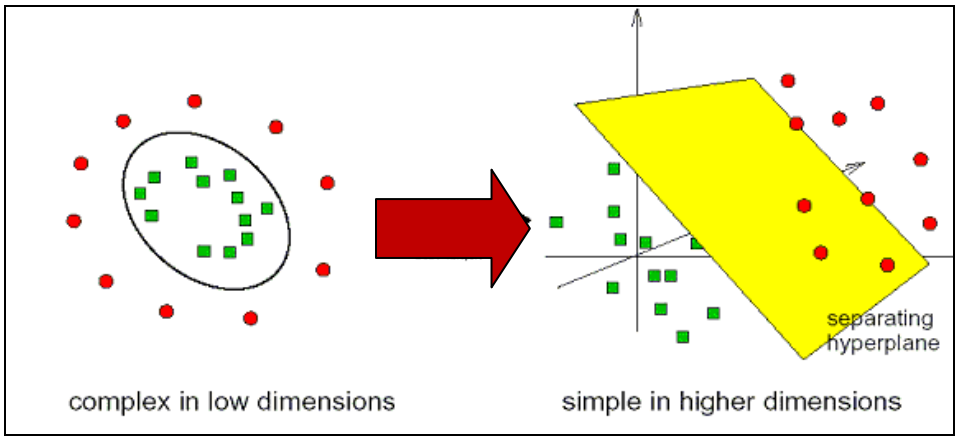


Figure 21. Visual representation of a complex classification scenario converted to linear discrimination at higher dimensions by support vector classification. (Source: Improved outcomes software website, 2004)

Perfect separation will not be always possible. The lagrangian parameters indicate the importance of a SV defining the hyperplane. The value can be extremely high if the SV is located in a zone where there are predominantly another data class, leading to a less reliable classification due to data overfitting (Zomer, 2004). It is solved, such as in the case of SVM regression, with the concept of soft margin classifier. With the addition of a regularization parameter in the constraints that controls the margin maximization ($0 \leq \alpha_I \leq C$) while the number of misclassified samples is kept inside an acceptable range the overfitting problem is solved.

SVM was originally designed to classify two classes, but there are few approaches that allow multiclass classification reducing the data to few binary classification problems. The “one against all” approach is based on developing classification of a single class from the rest of classes, and it is done for all the classes in the set. Each new sample to be classified acquires as many decision values (from decision functions) as data classes, and

it is classified in the class where the sample obtained the highest value (Hsu and Lin, 2002). One of the problems of this method is the errors that can be derived of using unbalanced number of classes. The ‘One against one” method seems to be better for multiclass classification purposes, but it is very memory demanding since it works with pair of samples, comparing the new sample to each sample from the data set and keeping the most predominant class according to the decision function. The winning class is assigned to the sample.

Recalling, ANN and SVM are powerful alternatives for identifying patterns or classifying data in higher dimensions (cannot be separated with a hyperplane in the actual dimensions). For more straight forward classification problems, some other methods with low complexity should perform well. K-nearest neighbor clustering is a simple yet successfully used alternative for supervised classification, and similarly K-means is a good simple alternative for unsupervised classification. Both methods are based on similarity among data based on distance measures and do not require much computational efforts or need of normality distribution of the variables. However, those do not have much popularity in NIRS. Other methods are based on developing PCA. An example is the Soft Independent Method of Class Analogy (SIMCA) and PCA followed by either linear or quadratic discrimination analysis are briefly discussed. These methods are briefly discussed in the following sections.

4.6.1. K-nearest neighbor

The similarity between two data points can be measured by either the distance between them (deterministic methods) or by probabilities of belonging to certain cluster (probabilistic classification) (Tran et al., 2005). K-nearest neighbor and K-means are deterministic methods.

In the K-nearest neighbor algorithm, usually the Euclidean distance of each sample to be classified to every other sample in the set is calculated. This way, K odd number of closest neighbors is chosen and the wining class labels the new sample. The optimal number of neighbors k may be calculated by iterations, although generally low values

from 1 to 5 tend to work the best. Despite the simple approach the method, the results obtained by k-nearest neighbor can be as good as the more complex methods such as ANN, but there is a need to introduce other criterions of majority if the number of samples belonging to the different classes is not similar (Massart et al., 1998).

4.6.2. K-means

K-means algorithm is similar to the k-nearest neighbor but it is an unsupervised classification method so previous class information is not required. It starts assigning the first k samples (where k is again user-selected) a cluster class. The next samples to be classified are assigned to any of the existent clusters according to the distance to the centroid of each cluster. Once the sample is assigned to a certain cluster, the mean centroid of the cluster is updated. A second check is performed, and the distance of each sample to the centroid of the existent clusters is calculated. If a sample initially assigned to a cluster is found to be closer to another cluster centroid, the sample is moved to the new cluster and both previous and current cluster means are updated. There usually are several iterations until convergence is achieved and there are no new assignments.

4.6.3. Discriminant Analysis

The classification is performed after creating an explicit optimal boundary between classes known *a priori*. When the boundary is created, either using a straight line (linear discriminant analysis, LDA) or a quadratic function (quadratic discriminant analysis QDA), all samples are classified in any of the existent classes. That can be a problem for samples that may belong to new masked classes or samples that belong to unspecified class because of being outliers.

For LDA, also known as Fisher's discriminant analysis (1963), and supposing the case of two classes (Figure 22), where the two ellipses that had been generated by the same confidence probability values touch, a line or plane tangent to the two ellipses can be drawn ($y = w^T x$) which maximizes the separation of classes. Another line or plane

perpendicular to the last one will serve as a projection line or plane, where the new value from each sample or score are defined by a linear combination of the initial variables. LDA maximizes the between-class scatter while minimizing the within-class scatter this way. We will be looking for a projection where examples from the same class are projected very close to each other and the projected means from each class are as far apart as possible. At the end, there will be as many linear equations as sample classes. The general equation for any class t has the form of equation 37.

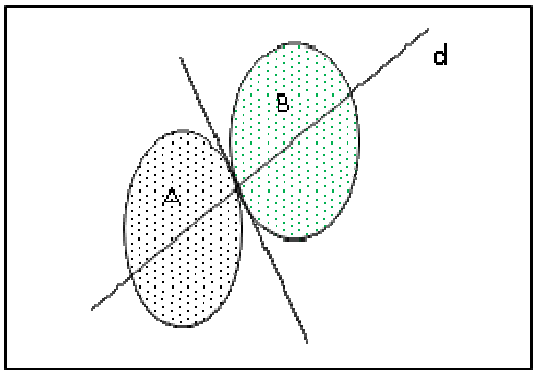


Figure 22. LDA for two classes and two variables.

Equation 37. $D_t = w_0 + w_{t1}*x_1 + w_{t2}*x_2 + \ldots. w_{tn}x_n = W'X + w_0$

D is the dependent variable, a discriminant score or membership value for class *t*. In NIRS, x would be wavelengths or PCs. *w* are the weight coefficients for each variable n that predict a class t (w_0=0 if data is standardized) which are calculated under the assumption that the variance of each class is the same. This is an important detail that should be taken in account since the pooled variance (Sd_{pooled}) is used for calculating the weight as shown in equation 38 for a two class classifier. Implicitly, LDA assumes that the mean from each group is the discriminating factor, not their variance.

Equation 38. $W' = (\bar{x}_1 - \bar{x}_2)' \cdot Sd^{-1}_{pooled}$

Where $\bar{x}_1$ and $\bar{x}_2$ are the averages of samples in group 1 and 2 respectively. The way that the projection line is created is following the direction that gives the optimum separation between classes thanks to the weight coefficients. The classification rule is defined for class A as D<0 and class B as D>0. In the case of more than two classes, the classification rule is based on choosing the highest value generated by a new sample (x_0) when plugged in equation 39 for each class t.

Equation 39. $$D_t = x_0^T S^{-1}_{pooled}\ \bar{x}_t - \frac{1}{2}\bar{x}_t^T S^{-1}_{pooled} \cdot \bar{x}_t + \log(p)$$

Where t is the class, $\bar{x}_t$ the average of values in that class, and the last term p is the prior probability of the group that would account for the different population size in each class. LDA works the best in situations that the centroids from the data classes are far apart forming dense data clusters: The variability between classes is high and the variability within class is small. Other of the assumptions that should be taken in account is the multivariate normal distribution of the samples belonging in each class.

With quadratic discriminant analysis (QDA) the groups or classes does not need to have the same variance since its discrimination function considers each individual class variance independently. This can be seen in the corresponding decision equation 40:

Equation 40. $$D_t = -\frac{1}{2}\log(|S_t|) - \frac{1}{2}(x_0 - \bar{x}_t)^T S_t^{-1}(x_0 - \bar{x}_t) + \log(p)$$

5.4) Soft Independent Modeling of Class Analogy (SIMCA)

SIMCA (soft independent modeling of class analogy) is a close alternative to PLSDA which is proven to perform when classes are not very homogeneous and are more spread in the projection plane and/or overlapped – this is also called “soft” classification

problem - (Wold et al., 1976). In fact, Frank and Lanteri (2001) reported the similarity of SIMCA to QDA. PCA is performed in each predefined class, where users have the chance to select the best number of PCs to be kept for each class (using for instance the eigenvalues). Choosing the number of PCs that should be kept is a critical step, enough PCs should contain necessary class information, but too many would diminish the signal information adding noise. When a new sample is presented to be classified, its residual variance after being fit in each class model is measured and compared with the average residual variance of a class, for instance using a F test. The upper limit would be set by the training samples already belonging to the class.

SIMCA has some advantages to PLSDA besides the one referred to classes homogeneity lately mentioned, such as the fact of being able to assign a sample to more than one class when falls between two classes with not clear assignment or not being assigned anywhere if the residuals exceed the limits from all classes, while PLSDA or K-NN only allow assigning samples to one single class, masking problematic classification cases (Kalivas, and Gemperline, 2006).

4.7. Variable Selection

Variable selection is a process that can be performed on both the principal components and on wavelength variables. In certain cases there may be previously available information from the sample to be analyzed and the wavelengths that may offer the most correlation with the analyte of interest. This could facilitate to choose the wavelength range of maximum information, although since NIR information is reproduced more than once in the whole NIR range, working with fewer wavelengths does not assure better calibration models for all applications. Other fast preliminary assessing techniques include the use of derivatives to find the peaks which height varies proportionally to the concentration of the compound to be measured.

Checking the values of the regression coefficients from PLS and PCR models give an idea of the wavelength region of interest and may bring the chance to improve the model removing the wavelength where the regression coefficients are not significantly different from zero or show noisy behavior. Similarly for ANN and as discussed, the input weights

of the hidden neurons can be checked to find the variables of more relevance, although high complexities of the net (many hidden neurons) makes this task more challenging. Basic methods of stepwise addition or stepwise variable elimination consist in starting the calibration with an initial number of variables and increase or decrease the number to be included in the model consecutively while monitoring the predictions (validation statistics). Validation stops once the results do not further improve. These methods are high time consuming, and stepwise addition may omit later variables that may contain valuable information.

While common methods for selecting the optimal number of cumulative PCs or latent variables in PLS and PCR regression have been previously discussed in the corresponding section, there are other more sophisticated methods that allow selecting a combination of individual PCs and wavelengths that lead to optimal regression or classification results. One of the most popular evolutionary algorithm is the genetic algorithm for variable selection. It is inspired by the natural genetic evolution: crossing of individuals, mutations, and the survival of the best individuals. The algorithm starts with a high number of random individuals, which are represented by their genome or group of genes – for our applications, the genes would be wavelengths -. The individuals are pictured as possible solutions or combination of genes. A fitness function is defined in order to asses how good or successful is an individual, such as the percent correct classification if the algorithm selects variables for classification, or RMSECV for calibration. By crossover, two individuals that survive to the selection act as parents to a generation of other individuals, interchanging some of their genomic information. The process of mutation imitates natural environments and randomly changes one or two digits in the individuals created by the crossover of two parent individuals, allowing keeping the population variability. The process stops after a fixed number of generations or when the goal is achieved (i.e. % misclassification).

4.8. Validation Procedure and Statistics

An adequate validation of the calibration models is a crucial step to determine the suitability of the model to predict new samples, which is the whole purpose of developing NIR calibrations. Ideally, the best validation should be done with distributed samples which were not previously used for calibrating. Since independent validation may not always be possible, cross-validation can provide a basic assessment regarding calibration performance. Because the final calibration model is not tested but rather several submodels developed with calibration data subsets, any statistic reported from cross-validation cannot be directly compared or interpreted the same way that statistics from a real validation of the final model with new samples. The standard errors from cross-validation are often optimistic and, especially in k-fold validation, highly affected by data artifacts (Naes et al., 2002e). However, reporting cross-validation statistics are preferred over reporting calibration results alone.

Table 1 shows the most used NIR validation statistics among the suggested and detailed in Williams (2001). However, it is not unusual to find literature using other statistics, reporting not so relevant figures of merit, or simply not reporting enough information for a good statistical assessment of the model quality. The coefficient of determination (R^2), which provides an estimation of how much variance between reference and predicted values is explained versus the total variance, seems to be one of the erroneously preferred guides for validation assessment. Its high dependency on the reference value range is often ignored (Fearn, 2002). The standard error of prediction (SEP, or SECV when reporting cross-validation results) provides information regarding calibration precision. SEP is corrected for the bias value (or systematic error); thus, when reporting SEP bias must be reported as well. The square root of mean standard error of prediction (RMSEP) is related to SEP and Bias according to equation 41. Because RMSEP accounts for bias and provides information regarding calibration accuracy, it can be reported alone, especially when bias is small (then RMSEP ~ SEP) (Davies and Fearn, 2006).

Equation 41. $$RMSEP^2 = SEP^2 + Bias^2$$

The final statistic to be discussed is the ratio of performance of deviation or relative predictive determinant (RPD), which is dimensionless and specific of NIR spectroscopy. It is related with the ability of the model to predict future data in relation to the initial variability of the calibration data. Basically, if a calibration leads to a low SEP but the calibration was carried out with a small range of reference values (standard deviation of reference values almost the same as SEP), the model would only be predicting the data average. Williams (2001) provides ranges of RPD values related to the calibration suitability: values above 8 indicate that the calibration can be used for any purpose, while values below 2.3 indicate a poor calibration performance, with use for predicting new samples not advisable.

Table 1. Table of common validation statistics for NIR calibrations

Statistic	Units	Equation
Coefficient of Determination (r^2)	Unitless	$r^2 = \frac{\left(\sum_{i=1}^{n} \hat{y}_i y - \sum_{i=1}^{n} \hat{y}_i \sum_{i=1}^{n} y_i \Big/ n\right)^2}{\left(\sum_{i=1}^{n} \hat{y}_i^2 - \left(\sum_{i=1}^{n} \hat{y}_i\right)^2 \Big/ n\right)\left(\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n\right)}$
Standard Error of Prediction (SEP)	Same as reference values	$SEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i - bias)^2}{n-1}}$
Root mean square of the error of prediction (RMSEP)	Same as reference values	$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$

Statistic	Units	Equation
Bias (d)	Same as reference values	$d = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)}{n}$
Ratio of performance of deviation (RPD)	Unitless	$RPD = \frac{Sd_y}{SEP}$

$\hat{y}_i$= ith validation sample predicted value

y_i= ith validation sample reference value

n=number of samples in validation set

Sd_y = standard deviation of reference values from the validation set

5. Near Infrared Spectroscopy for Single Seed Analyses

5.1 NIRS applications in quantitative analysis of single seeds

One of the limitations in single seed analysis is the small sample size and thus the amount of the compound to be measured, which may be below the NIR detection limits. Transmittance measurements seem to be more attractive for measuring low concentrations since the light goes through the whole sample, but seeds are variable in size and this negatively impacts the light pathlength (Cogdill et al., 2004). Low repeatability of spectral measurements in function of kernel mass/size have been reported (Delwiche, 1995), which seemed to be weaker in reflectance mode (Delwiche, 1998). Reflectance measurements may not be so affected by the seed pathlength, but they are affected by sample heterogeneity. This is important for grains such as corn, which have a differentiated germ side and the side to be analyzed matters. Janni et al. (2008) indicated that kernels are opaque to wavelengths above 1300 nm, which are common in reflectance

measurements, depending on their size. Furthermore, it has been indicated that higher wavelengths may not produce linear responses between spectra and analyte concentration in single kernel analysis (Delwiche, 1998).

There are several conventional chemical methods for single seed analysis which have been applied and correlated with NIRS spectra in the literature. Those are applied to measure some major compounds such as moisture, protein, and oil. NIRS is dependent on an external reference method, and for this reason the resulting calibration precision can be as maximum as good as the reference method. It has to be considered that other errors specific from a given laboratory are added to the intrinsic error associated with the method.

Some methods are already very popular in single seed analysis. Pulsed low resolution NMR is preferred for absolute oil analysis (Alexander et al., 1967). This method gives more precise results than Soxhlet or supercritical fluid extraction, as pointed by Robertson and Windham (1981) and Cogdill et al. (2004), who targeted the reference method as one of the sources of error for their corn kernel oil calibration. Moisture can be determined by either mass difference after seeds are oven-dried and low resolution NMR, simultaneously to oil determination. Protein is usually determined by combustion, giving precisions around 0.2% (Hunt et al., 1977). Protein determination by Biuret, a colorimetric assay, is reported to be more accurate for larger samples (Baianu et al., 2004).

Other minoritary compounds such as fatty acids and amino acids are found in the seed at very small concentrations, and single kernels themselves suppose a very small sample amount to be analyzed so the resulting NIRS calibrations are expected to have average correlations between reference values and spectra with high prediction errors.

Reference values or determined compound can be expressed as absolute value (i.e. mg/kernel) or relative units (weight percentage of the compound on seed weight). Different moisture basis can be used to express the seed weight: (1) current moisture of the seed at the time of analysis (as is), (2) percentage of the compound on dry seed weight, or (3) percentage of the compound on seed weight at a given moisture level. That

last option is common for soybeans (usually 13% moisture level) and corn (15% moisture level).

5.1.1 Analysis of grains

Corn. Corn kernels have been the grain more extensively analyzed by NIR single kernel quantification due to its significance in US, its complications due to the heterogeneity in physical traits, and the extensive breeding programs to achieve highest oil and starch concentrations. The first NIR quantitative analysis in corn single kernels is dated back in 1978 for determining moisture. Finney and Norris (1978) used one of the first NIR transmittance monochromators instruments (Cary) and a multiple linear regression (MLR) approach on a short wave range (920-950 nm). Their results gave a very good correlation spectra-moisture content (R^2= 0.93) and accuracies of 2 – 3% by weight.
Quantification of oil has been of high interest as reflected by the number of studies that can be found in the literature. Corn oil is of high value in the area of biodiesels and food, and research is focused on developing corn varieties with high oil content in corn while maintaining high yield in the field. The main results from the literature are summarized in table 2. Comparing those results from the literature is difficult not only due to different statistics reported (accuracy vs precision, correlations), reference methods involved in the analysis, or error contribution of the reference laboratories; but also the oil range in the calibration set and its effects on corn kernel physiologic structure according to Janni et al. (2008). Their reported results in table 2 considered samples up to 8% in oil content, and observed that working with samples at higher concentrations the errors increased considerably.

Table 2. Summary of NIRS oil calibrations for single corn kernel found in the literature. As is moisture basis used, best preprocessing method applied to the spectra.

Technology	Validation Samples	R^2	RMSEP (%)	SEP (%)	Key Reference
Conventional Transmittance	73	0.75		** SECV 1.20[(b)]	Orman and Schumann, 1992
Imaging Transmittance	151	0.54		** SECV 1.38[(b)]	Cogill et al., 2004
Imaging Reflectance	100	0.67	1.10 [b]		Weinstock et al., 2006
Tumbling Kernel Reflectance	80	0.94	1.5 - 0.40 [a,b]		Janni et al., 2008
Light Tube Reflectance	115	0.86	0.79 [b]		Spielbauer et al., 2009
Rotating Cup Reflectance	~250	0.95	0.60 [c]		Jiang et al. 2007

a. Oil concentrations in calibration below 7.8%, 12 seconds of analysis.
b. As is weight moisture basis
c. Dry weight moisture basis

The earliest oil predictions were carried out by Orman and Schumann in 1992 using transmittance measurements. Principal component regression (PCR), and mathematical preprocessing techniques of the spectra such as derivatives were applied. They performed multiple scans on the kernels, removing and replacing them and averaging the spectra in order to improve the signal-to-noise ratio, but they did not find improvement in the results probably because the effect of changes in kernel position overcame any improvement that could result from averaging several spectra. Cogdill et al. (2004) also reported the difficulties of sampling in transmittance measurements, although kernel positioning (germ up or down) did not affect in repeatability or spectra quality. Sample pathlength differences could have been one of the big error sources which impacted his study with NIR transmittance imaging. Each kernel was represented by 60,000 pixels,

and he averaged them to obtain a single spectrum per kernel previous to carry out the calibrations. His PLS calibration for kernel moisture had a lower coefficient of determination compared to the one reported by Finney and Norris (1978) (R^2=0.87) and cross-validation precision (SECV) of 1.04%. His oil calibration in as is moisture basis (table 2) gave lower accuracies than the ones reported by conventional transmittance, and a big part of that was attributed to the reference method, by solvent extraction, which is less precise than NMR used by Orman and Schumann (1992).

As previously mentioned, even though the side of corn kernel facing the light does not matter for transmittance measurements since the light goes through the whole seed, it does for reflectance measurements. A small portion of light penetrates the sample and gets to be measured as diffuse reflectance, so kernel size affects NIR reflectance spectra as noticed by Orman and Schumann (1992) in their PLS oil calibrations. Larger kernels lead to positive residuals in their models, thus oil was over predicted for those kernels. And it was the opposite for smaller kernels. The most recent quantitative analysis of corn kernels have been carried out by reflectance mode and it has been found the use of mathematical preprocessing methods such as standard normal variate (SNV), multiplicative scatter correction (MSC), derivatives and the combination of those to lead to calibrations with better predictive ability. When carrying calibrations with absolute (mg/kernel) and relative units (%), the first gave higher coefficients of determination (R^2) and better predictive ability in terms of RPD, but were negatively affected by SNV and MSC preprocessing (Spielbauer et al, 2009). That could suggest that those pretreatments remove information related to seed size or information relevant for determining the absolute composition.

Jiang et al. (2007) worked in conventional reflectance with a rotating cup and developed PLS calibrations with over 600 kernels for oil, protein, and starch. Kernels were scanned 10 times both germ up (facing the light and detector) and down, and developed calibrations individually for each side and averaging both up and down spectra. Averaging both sides lead to the best predictive models for protein (R^2=0.95, RMSEP=0.30%). Germ-up spectra worked best for starch calibrations (R^2=0.89, RMSEP=0.96%) while germ-down, curiously, worked the best for oil prediction reported

(table 2). By reflectance imaging, Weinstock et al. (2006) quantified oil and oleic acid isolating the kernel germ. Spectra from all pixels representing the germ were averaged for their calibrations. The best predictive calibrations were obtained using genetic algorithms for optimal wavelength selection and PLS, after preprocessing the data with first derivatives. With 7-8 optimum wavelengths selected, oil predictions were less precise than the ones reported by Jiang et al. (2007) (table 2), but still better than transmittance calibrations. The prediction of oleic acid was less successful, as expected due to the low concentration of this compound in individual kernels (R^2=0.65, RMEP=13.7%). They also found that for their static imaging unit, kernel orthogonal orientations (i.e. 0 degrees vs 90 degrees) on the focal plane had as much as impact as positioning the corn kernel upside or downside, in terms of spectra baseline effects. Fortunately, that could be eliminated by baseline subtraction from the spectra. For other applications such as discrimination of wheat infestation discussed later, the kernel orientation in imaging units did not show any difference (Singh et al., 2009).

It has been suggested that the predictive ability of calibrations developed by reflectance could be enhanced by total and homogeneous illumination, and measurements from the whole kernel in each side (Janni et al. 2008). Two recent approaches for kernel measurements by NIR reflectance have pursued this objective. Janni et al (2008) introduced an air-tumbling kernel approach, patented by Pioneer Hi-Breed international (Wright, S., 2007; Janni, 2007). They used a diode-array instrument with a pro, a pipette with air flow, and a duraflect coated tube to host the sample. The instrument scans the seed while tumbling for a certain amount of time. At least 12 seconds of analysis were required for good performance, according to optimization results. The resulting oil calibration was more predictive than the one reported for the rotating cup (table 2). This was in agreement by their validation of the tumbling kernel equation by rotating and static (both germ up and down) kernel spectra. Validation of the tumbling equation by averaging both sides spectra gave better oil predictions (RMSEP=1.60%, as is moisture basis) than validation with one side spectra, demonstrating that the tumbling technique was scanning the whole seed. The worse validation results were obtained from the one-side static cup (RMSEP=1.20%).

The problem of the tumbling approach is the long analysis time, which makes it not practical for continuous data collection. The second approach is proposed by Armstrong (2006), and it is based on a light tube illuminated by 48 tungsten lamps. A bifurcated fiber optic, one output in each extreme of the light tube, captures the spectra which are read by the spectrograph. The scanning time is trigged by sensors. The seed enters the tube where it is completely illuminated and the spectrum from a big part of the seed is collected. A single spectrum measurement takes 30ms. The tube is not wide enough to let the corn kernel tumble unless is a popcorn variety, thus the orientation of the corn kernel had a slight effect. Based on PLS moisture calibrations, the best precision was achieved introducing the kernel with dent end orientation in the tube (SECV=0.76%, R^2=0.97, RPD=5.5). A later study used the same instrument to develop PLS calibrations for starch, protein, and oil in corn kernels at dry weight basis (Spielbauer et al, 2009). Starch predictions showed the highest standard error of repeatability after taking multiple scans for a same kernel, highest SEP (3.72% and 18.2 mg/kernel) and lower coefficient of determination (R^2=0.66 and R^2=0.85 for relative and absolute units, respectively). Starch absolute values were strongly correlated with the seed weight, and it was seen that linear regression between seed weight and absolute starch content leaded to better predictions than the NIR calibration. Kernel weight was highly correlated with the NIR spectra (R^2=0.86). The best protein calibration was similar to the oil calibration in table 2 (R^2=0.88, SEP=0.81% and SEP= 3.82mg/kernel in as is moisture basis).

Several researches use the bulk reference data for single corn kernel calibrations to avoid the expensive reference analysis of single kernels. Finney and Norris (1978) realized how averaging the individual kernel spectra, the precisions were better for predicting a whole bulk sample moisture than the precisions obtained for single kernel prediction. Tallada et al. (2009) worked with the light tube from Armstrong's (2006). Averaging kernel spectra and using bulk reference data, they developed PLS calibrations for tryptophan, lysine, protein, oil, and soluble sugars. Calibrations using relative units (% dry basis) did not give good predictive models, with R^2<0.5 and RPD below 2 for most of compounds. Absolute value units (mg/kernel), similar to previous studies, gave models with predictive abilities suitable for gross screening when no spectra preprocessing was

applied, with precisions very similar to the ones they achieved using single kernel references. For protein, RPD=3.15, R^2=0.89, and SEP=3.6 mg/kernel. For Oil: RPD=2.53, R^2=0.88, and SEP=2.49 mg/kernel. And mass: RPD=2.98, R^2=0.88, and SEP=0.02 mg/kernel.

Wheat. Determination of protein content and hardness in wheat is important because both attributes affect the quality of flour. The first protein PLS calibrations for single wheat kernel reported used transmittance (Delwiche, 1995). Different wheat classes were involved (white, red, soft, and hard) and independent calibrations for each class were developed. Kernels were only scanned once (no multiple spectra averaged) since the old instruments in the study took long time (about 3 minutes per scan). The best results with second derivative and MSC spectra preprocessing were promising: R^2 > 0.85, SEP between 0.4 and 0.9 % (12% moisture basis), and bias values up to 1% depending on the analyzed class. Similar results were later achieved by Nielsen et al. (2003), with RMSEP=0.48% (dry basis). By reflectance measurements, Delwiche (1998) showed that developing a general calibration with all what classes did not reduce prediction accuracies. Testing both PLS and MLR models with up to 8 wavelengths, the best predictions were achieved with PLS: SEP values ranged 0.42 to 0.59%, R^2 of 0.90 – 0.97, RPD between 2.88 and 4.72, and absolute biases much lower than the previous transmittance studies (below 0.2%). A later study carried out by Delwiche and Hruschka, (2000) averaged single kernel reflectance spectra, from 10 to 100 kernels and each single kernel spectrum being the result of averaging 32 scans or spectra, in the short wavelength region from 1100 to 1398 nm. The difference in cross validation precision decreased fast when averaging spectra from 10 to 30 kernels (SECV from 0.35% to 0.22%, 12% moisture basis) and continued decreasing at smaller intervals when averaging 50 and 100 kernels (0.18 and 0.16%, respectively).

An attempt to predict vitriousness was carried out by Nielsen et al. (2003), who correlated single wheat kernel transmittance spectra with red reflectance measurements by RGB image analysis. Validation statistics were not acceptable (R^2=0.58 and RMSEP= 4.6 absorbance units). Not successful predictions were obtained for neither kernel density

employing diverse reference methods nor kernel hardness. Delwiche (1993) also developed transmittance hardness calibration using reflectance predictions using the American Cereal Chemists method (1983) for obtaining references. R^2 equal to 0.70, SEP about 15 hardness units, and Bias around 1 harness units were obtained. He stated the need of improvement of the reference method and the existence of other factors which affect grain hardness and NIR cannot measure. Better results were achieved combining visible light and NIR reflectance, but after averaging from 5 to 50 kernels from each class (Maghirang and Dowell, 2003). Averaging 50 kernels gave the best results (R^2 = 0.91, SECV=7.57 hardness units, and RPD=3.35).

Dowell et al. (1999) carried out a study to predict mold damage by scab (*Triticum aestivum* L.), vomitoxin, and ergosterol (alcohol associated with the mold invasion) in wheat kernels. Vomnitoxin calibration was developed for levels above 5 ppm, since below that limit the performance was considerably worse, probably because of working far below the NIR detection limits. The results, although there was some correlation between the spectra and the variables, were not satisfactory: R^2 = 0.66 and SEP = 52 ppm for vomitoxin, and R^2 = 0.64 and SEP = 108 ppm for ergosterol.

5.1.2 Analysis of oil seeds

Oil is not the only compound of interest to be quantified in single oilseeds and several researches carry out the quantification of other specific compounds. For instance, the first study analyzing single soybeans determined moisture by transmittance (Lamb and Hurburgh, 1991), PLS models lead to SEP values of 0.65-0.69%. Later, Armstrong (2006) used the light tube proprietary instrument to determine moisture in soybean seeds by reflectance mode, and obtained SECV=0.32%, R^2=0.99, and RPD=10.5. He also developed PLS protein calibrations (SECV=0.99% at 13% moisture weight basis, R^2=0.96, and RPD=4.9). He found MSC to be the preprocessing method that leaded to more precise predictions. Those results were on the same order of Tajuddin et al. (2002) who worked by transmittance (SEP=1.32% - 1.57%, dry weight basis) and used Biuret as reference method instead of the common combustion. Calibrations with highest

precisions were obtained from soybeans with higher diameters (>6 mm). An oldest research carried by Abe et al. (1996) reported protein calibrations with better prediction statistics (SEP=0.67%) when analyzed spectra was the average from two measurement points, in transmittance mode. Baianu et al. (2004) also reported promising results using both a diode array reflectance instrument and a Fourier Transform (FT). SECV was 0.3% for the FT-NIR instrument and 1.1% for the diode-array (dry weight basis). Delwiche et al. (2006) developed protein and inorganic phosphorous calibrations for single seed soybeans using absolute reference values (g/kg for protein and mg/kg for phosphorous). They achieved calibrations with low predictive ability: RPD=1.2 for both protein and inorganic phosphorous, R^2<0.50, and RMSEP = 13.93 g/kg for protein and 568.6 mg/kg for inorganic phosphorous. Protein models have been also developed for rapeseeds in the literature using reflectance mode. First, Velasco and Mollers (2002) obtained calibrations that leaded to R^2=0.94 and SEP=0.77%. Hom et al. (2007) obtained similar results with cross-validation (R^2=0.96 and SECV=0.74%, dry weight).

Tajuddin et al. (2002) also developed PLS oil calibrations for soybean seeds with the same grating reflectance unit used for protein calibrations, obtaining SEP= -0.14% and -0.09% (as is moisture weight basis) for smaller (< 6 mm) and larger seeds, respectively, using hexane and chloroform extraction as reference method. Calibrations using the same reference method were developed by Baianu et al. (2004), and SECV were on the same range (SECV=0.2% for FT-NIR and 0.5% for diode-array reflectance).

For husked sunflowers achenes, Sato et al. (1995) carried out a research examining the correlation of fatty acids extracted by gas chromatography and reflectance NIR spectra. Errors were not reported, but correlations were above 0.90 for linoleic acid and total fatty acids. Velasco et al. (2004) obtained promising PLS reflectance models for linoleic acid (R^2 = 0.93, SEP=64.3g/kg), oleic acid (R^2=0.92, SEP=82.9g/kg), and stearic acid (R^2=0.83, SEP=42.0 g/kg) in husked achenes of sunflowers, as is moisture weight basis. Previously, Velasco et al. (1999a) carried out some work to determine oleic and linoleic acid in unhusked achenes, obtaining R^2=0.88 but predictive errors very close to the husked achene calibrations. It was concluded that removing the husk from sunflower achenes did not improve the calibration performance for those two fatty acids. Another

research carried out by Tillman et al. (2006) was based on developing oleic and linolenic acid PLS calibrations in single peanuts using FT-NIR. The results from those calibrations were better than the ones achieved for sunflower seeds, probably because of the bigger seed size (RMSEP=20 kg/g for oleic acid and RMSEP=19 g/kg for linoleic).Rapeseeds also showed significant correlation coefficient between NIR spectra and their extracted oil (R=0.86) (Sato et al., 1998). Hom et al (2007) predicted oil content using the gravimetric extraction method for obtaining seed references, and the calibrations were promising (SECV=1.14% in % dry seed R^2= 0.97). Individual fatty acid NIR calibrations could be also developed. Using relative units respect to the percentage of total fatty acids, Niewitetzki et al. (2010) obtained SEPs=2.7 – 3.7% for oleic acid (R^2 up to 0.91), SEP=1.2 – 1.8% for linolenic acid (R^2 up to 0.90), and SEP = 2.5 – 4.2 % for linoleic acid (R^2 up to 0.78). Hom et al (2007) also analyzed other compounds which leaded to calibrations with rough screening abilities. Those were total glucosinates (SECV=10.3µmol/g dry weight, R^2=0.86), alkenlyl (SECV=9.29 µmol/g dry weight, R^2=0.83), and Indole (SECV= 1.35 µmol/g dry weight, R^2=0.86). One of the calibrations, total aromatic compounds, was not successful (SECV=0.34 µmol/g dry weight, R^2=0.36).

5.1.2 Analysis of other seeds

Researches concerning the quantification of organic compounds from other seeds can be found in the literature. One of the oldest researches was carried out by Patrick and Jolliff (1997) regarding the use of transmittance for quantification of meadow-foam seed oil. A couple of calibrations were developed using NMR as reference method: SECV of 3.6 and 4.4%, and R^2=0.95. Those results were compared with the ones that were achieved for corn kernel oil calibrations with similar total oil range, although problems with seeds with oil content under 5 mg were reported. Armstrong's light tube approach (2006) was used to develop calibrations for common beans (*phaseolus vulgaris L.*) for protein, starch and seed weight (Hacisalihoglu et al., 2010). Protein PLS calibration in dry weight basis gave SEP=1.6% and R^2=0.82, SEP=4.9% and R^2=0.56 for starch, and SEP=41.2 mg and R^2=0.74 for seed weight. Simmilarly to soybean calibrations, starch showed the highest

difficulty to be calibrated. Protein calibrations could be compared to Armstrong's for soybean seeds, which validation was not carried out with an independent set but by cross-validation, thus expected SEP values would be higher.

5.2 NIRS for seed sorting and discrimination

Discriminative studies intend to set or classify seeds in two or more groups. Besides using qualitative attributes for classification (i.e. infested vs non-infested seeds), any quantitative calibrations which shows prediction errors valid for rough screening (RPDs around 2-3) can be thought as a two-class discriminative model: high analyte content vs low analyte content. This gives the chance for calibrations with low accuracies/precision to be used with the underlying interpretation of a discriminative model and still be usable to create more homogeneous populations. With the currently improved grain varieties developed following specific applications, sorting seeds to increase purity and homogeneity in batches is of high interest at farmer's and industrial levels. From a breeder's point of view, most of the times the interest is in finding the seeds with the highest or lowest concentration of certain compound, without focusing on the exact amount. For instance, Orman and Schumann (1992) determined that their oil prediction for corn kernels resulted enough for screening and they wanted to use it as a method to select the kernels with highest oil content. They compared their NIR prediction results with their reference method, NMR, and they found out that from the top 25% corn kernels with highest oil concentration only 58% of them were in agreement with the top kernels selected from NMR results. Interestingly, scanning each kernel 10 times and averaging them, resulted in 75% of agreement even if this did not translate in better prediction error in the quantitative analysis.

5.2.1 Grain

Damage, toxins and infestation are popular research topics in NIR discriminative analysis of grain. Toxin contaminated corn kernel and wheat infested kernels have been widely

analyzed by NIRS. Furthermore, common pest insects were studied by NIRS without being associated with the grain. Dowell et al. (1999) could differentiate primary and secondary pests insects over 99% accuracy using neural network model. ANN and PLS-DA could get accuracies over 95% in classifying the insects within a genus. Dowell et al. (2000) worked discriminating fly puparia and its parasitoids, and the PLS-DA the classification accuracies ranged from 80 to 90%. Even the discrimination among insect types and sex is possible. Tsetse fly pupae sex could be determined by reflectance (950 – 1700 nm) and PLS-DA, with accuracies up to 97% for pupae from 1-6 days previous to emergence. The accuracy dropped to up to 75% for pupae of 23 days before emergence (Dowell et al., 2005).

Innovative and current on-going researches involve discrimination of transgenic and conventional grains. In a recent study of Jiao et al. (2010) transgenic rice was discriminated combining spectroscopy and conventional chemical methods that proved that the differences were mainly due to protein, aminoacids, two fatty acids, and two vitamins.

Corn. Few researches concern the discrimination of corn kernel according the differences in their endosperm characteristics such as vitreosity, floury, and hardness. The difference in varieties of single and double-mutant recessive alleles which affect starch structure in corn kernels could be discriminate by Campbell aet al. (2000). The normal corn kernels vs mutants could be classified with PLS-DA with high accuracy as the discrimination was possible also visually most of the times, leading no misclassifications when using 9 latent variables. The discrimination accuracy between mutants from a same class was not constant and depended on the variety. They also worked on discriminating low and high amylose content kernels, which lead to 70% of correct classified samples in the low concentration group, and 90% for kernels belonging to the high concentration class. Manley et al. (2009) used reflectance chemical imaging to analyze kernel endosperm to differentiate flourly and vitreous endosperms. The PLS-DA model was acceptable (R^2>85%) and they found a third class of endosperm which show characteristics from both flourly and vitreous endosperm and which was visually differentiable in the PCA

score plots. Williams et al. (2009) classified single kernel endosperm spectra in either glassy or floury at 99% accuracy. Hardness, which besides being a genotypic expression is also affected by environmental and handling factors, could also be deduced in this study from the ratio of glassy and floury endosperm of each kernel.

The raising problem, especially in harvest years with high precipitation levels, of fungal toxin contamination of corn and the associated risks in human health, lead to find for methods of fast screening of corn batches. Single kernel analysis is very adequate, since representative sampling in bulk batches is difficult due to the fact that toxins may be found in only few highly contaminated kernels (Pearson et al., 2001). Detection and classification of corn kernels with toxins was reported by Perason et al. (2001) for aflatoxins (toxin from *Aspergillus flavus* fungi) and by Dowell et al. (2002) for fumonisin (toxin from *Fusarium verticillioides* fungi), both using reflectance and transmittance NIR scanning modes. For both studies, discriminant analysis (DA) using few wavelengths and PLS discrimination analysis (PLS-DA) for the entire range were tested as discriminative algorithms. For discriminant analysis, three classifications based on toxin concentration thresholds were created: 1ppb, 10 ppb and 100 ppb for aflatoxins; and 10, 50 and 100 ppm for fumonisins. The best threshold for toxin detection based on correct classified kernels (positive or negative) was 10 ppb for aflatoxin, and 10 ppm for fumonisin. In both studies, DA performed better than PLS-DA, and was suspected that transmittance could lead to more accurate results. This could be especially true in the case of aflatoxins, since the mold may stay inside the kernel. For aflatoxins also, the germen facing the detector in reflectance provided better classifications, but the kernel position was not relevant in the fumonisin study. The best aflatoxin classification with DA misclassified over half of the kernels with toxin content between 10 and 100 ppb, but 95% of the kernels out of that range (either lower or higher concentrations) were correctly classified. This pattern was slightly better for fumonisin, where the best classification lead to 23% of the kernels between 10 and 100 ppm misclassified. Although those studies showed that NIR can discriminate between low contamination levels and very high, the application cannot be used for safety control.

Wheat. Similarly to corn researches, wheat endosperm has been also analyzed by NIR. Vitreousness is an indicator of quality as it is related to protein content and an attribute that leads to better qualities in foods like pasta and other derived products. Dowell (2000) proved that NIRS could discriminate vitreous and non-vitreous single kernels of wheat, and the best accuracy (99%) was achieved with kernels that were clearly distinguishable by inspectors as on of both classes. Some kernels were not clearly one class or the other, and when introduced in the calibration made the accuracies drop to 75%. In his study, he concluded that protein or starch concentrations together with light scattering effects could be what NIRS was measuring. On the other hand, In Manley et al. (2009), corn kernel virtuousness and floury classifications showed the highest beta coefficients in the carbohydrate and water region – meaning that what NIRS was measuring for discrimination was mainly starch and/or water binding-. Wang et al. (2002) carried out another study of classification of vitreous and non-vitreous spring wheat kernels, including defective kernels such as bleached, cracked and sprouted. He concluded that scattering was also a major contributor to the classification together with color, hardness, starch content, and protein concentration in agreement with Manley et al. (2009) and Wang et al. (2002). He achieved accuracies over 90% including defective kernels, but 75% of the bleached kernels were misclassified.

A couple of researches carried practical studies with NIR instrumentation and sorting devices to homogenize batches with mixtures of wheat. The discrimination of wheat kernels with high protein (>12.5% at 12% moisture weight basis) and low protein (<11.5%, 12% m.b.) was carried out by Pasikatan and Dowell (2004) in reflectance mode. It was estimated that with maximum two consecutive sorting processed, blends of 95:5 could lead to protein concentration (measured in two subsamples) of the initially dominant class of protein. Color and vitreousness seemed to drive the classification, thus reassuring that protein and vitreousness are both factors for quality assessment and can be detected by NIRS. This observation kept on agreeing with the previous results of discrimination of vitreous wheat kernels. Dowell et al. (2006) sorted wheat kernel from 4 mixture bins, being able to achieve at the end 4 bins with 1% increasing protein fractions, being the average difference of bins with high protein bins and low protein of 3.1%

points. They also classified the kernels by hardness, and NIRs could narrow the hardness distribution in each bin to 17 hardness units of difference between the highest hardness fraction and the lowest.

Discrimination among varieties is another way to improve the quality of a bulk batch of grain as different classes may be suitable for different purposes and have different end value. The mixture of classes may be sometimes accidental, by inadequate cleaning and handling, and this would result in a batch of mixed classes that have a lower value overall. Delwiche and Massie (1996) developed PLS-DA and MLR binary decision models to discriminate among wheat classes at screening level but with good repeatability. Classes that had different colors such hard white and hard red winter leaded to excellent accuracies around 99% using wavelengths close to the visible, and dropping to 78-91% when using the near infrared region. The tree decision technique later used for further classification of classes of the same color was not sufficient when only NIR data was used. In a later research, red and white kernels were classified using NIR reflectance and visible light with PLS-DA, attaining classification accuracies over 95% (Dowell, 1998). Waxy varieties, which are characterized for having starch which lacks amylose, were discriminated from partially-waxy and wild wheat varieties according to variation of amylose content by reflectance PCA, using discriminant analysis (Delwiche and Graybosch, 2001). The first PC allowed already 50% correctly classified kernels. Applying either linear discriminant analysis or quadratic discriminant analysis, maximum accuracies were around 70%. NIR seemed to detect the differences in amylose and amylopectin in the kernels, but the overlap in amylose content in classes could be what prevented to get better accuracies. Delwiche et al. (2006b) worked with durum wheat kernels, classifying them according to their possible 4 waxy alleles with PCA linear discriminant analysis (LDA). The full waxy genotype was classified with accuracies above 95% but the classification accuracy of the non-waxy genotypes was not possible. Dowell et al. (2006) besides classifying wheat, they also segregated waxy kernels for batch homogenization. They could increase the percentage of 94% of waxy millets in unsorted samples to 98% in 42 of the 48 samples, the 6 samples left had a decrease in waxy kernels (from 94.5 to 93%) probably due to reference method errors and lack of

representatibity of the selected kernels. Overall classification of wheat kernels from different Canadian varieties using NIR chemical imaging have been reported, with classification accuracies over 90% using LDA, QDA, and ANN classifiers (Mahesh et al., 2008).

Another NIRS single seed application which had big impact is the sorting of damaged wheat kernels, especially insect-damaged and infested kernels as it is a common problem in grain-storing facilities. Some insect species are easily removable by cleaning operations but other insects grow inside the kernels, making them invisible for methods based on visual inspections (Perez-Mendoza et al., 2004). By manual sieving and exhausting visual inspection, it cannot be detected insect concentrations below 5 insects per kilogram of grain (Wilkin and Fleurat-Lessard, 1990), so automated methods to help in this task is of high interest. The first NIR single kernel infestation researches involved the discrimination of wheat infestation by rice weevils (*Sitophilus oryzae).* Ghaedian and Wehling (1997) worked with PCA and Mahalanobis distances over full and partianl reflectance spectra range (1100–2498 nm and 1100–1900 nm, respectively) to discriminate between not-infested and rice weevils infested wheat kernels. The short wavelength region provided the highest accuracies and found that the region 1980 - 2498 nm had not usable information for the discrimination. Ridgway and Chambers (1998) worked with imaging reflectance and could get the best differences among sound and infested kernels substracting the images at 1300 nm from the images at 1202 nm. Not infested kernels appeared darker, which could be mainly due to loss of starch in the kernel. They later developed two models using short wavelength ranges (either 982-1014 nm or 972-1032 nm) which achieved accuracies of rice weevils larvae-infested vs not-infested wheat kernels over 96%. They agreed with Ghaedian and Wehling (1997) in the matter of not needing wide wavelength ranges for obtaining good discriminations. Dowell et al. (1999) PLS model for detection of scab-damaged versus absence of vomitoxin or ergosterol in wheat kernels also gave better results than the identification of scab –damaged grains by visual inspection. A research carried out by Baker et al. (1999) involved single kernel infestation by rice weevils and their associated parasitoid *Anisopteromalus calandrae,* which sometimes rears the host rice weevils in the

infestation. Kernels with pupae, larvae, parasite pupae, parasite larvae, and not infested kernels were scanned by transmittance. PLS-DA lead to classifications over 95% for infested vs not infested kernels, but the discrimination between infested with weevil larvae and parasitized weevil larvae was not possible. Maghirang et al. (2003) created an automated system to discriminate non-infested wheat kernels, infested kernels with dead rice weevils, and infested kernels with larvae/pupae at different growth stages. Live pupae got the highest accuracy (94%) while the small larvae got the lowest (63%). PLS-DA model created with pupae and large larvae when validated with the data from the same kernels stored through time (1 to 56 days) could be used to detect the insect, either dead or alive, with accuracies from 86 to 96%. The study of the damage to the kernels by insects using chemical imaging was investigated by Singh et al. (2009, 2010). Images from 1,101.69 and 1305 nm in form of a single PC scores, in agreement with Ridgway and Chambers (1998) results, were used for discrimination. LDA and QDA applied to several image features and statistics lead to accuracies ranging from 85 to 100%, depending on the insect causing the damage. When working with a colour imaging system, QDA seemed to work the best, correctly identifying over 96% of the healthy kernels and from 91 to 100% of the insect damaged (Singh et al., 2010). Overall, it was determined that the relevant wavelengths in insect infestation were the ones related to water from metabolic processes of insects, protein, lipids, phenolic compounds, and carbohydrates due to chitin insect cuticle absorption and a decrease of starch levels in the grain (Ghaedian and Wehling, 1997; Baker et al. 1999; Dowell et al., 2000).Another kind of damage, heat-damaged wheat kernels, was studied by Wang et al. (2001). The NIR region worked better than visible, and all kernels were correctly classified by PLS-DA. Light scattering was suggested to be a factor driving the classification.

Finally, the detection of sprouting of wheat could be assessed by reflectance chemical imaging in early stages by reflectance chemical imaging by Smail et al. (2006). The kernels, placed germ up, showed in the NIR absorbance image a wider area occupied by the germ when naked eyes could not see it. They concluded that NIRS could anticipate the detection of sprouting wheat much earlier than human eyes.

5.2.2 Oil seeds

Current NIR researches involving discrimination of oil seeds are mostly carried on soybeans. Wang et al. (2002) classified soybean seeds according to different kinds of damage: sound, heat, weather, mold, sprout, and frost. ANN with no hidden layer and taking VIS + NIR reflectance region (490 – 1700 nm) lead to good classification accuracies: 100% for weather, 98% for frost, 97% fr sprout, 64% for heat, and 97% for mold. In a later study, fungal-damaged soybean seeds could be again classified at high accuracies (99%) with PLS-DA and reflectance mode, which was also proven in the previus study, and seeds could be classified as sound or damaged by any 4 fungi (5 class model) using ANN at maximum overall accuracy of 94.6% (Wang et al., 2003). Kusamat et al (1997) could classify artificially aged soybeans from normal. DA using 2PC (PC2 and PC3) had a 60% of accuracy hen aged seeds were 3 days, 80% for 5-day aged seeds, and 100% when seeds went in aging procedure for 7 days. A decrease in phospholipid content in soybean seeds while ageing showed to be one of the main factors detected by NIRS.

5.2.3 Other seeds

Seed infestation, filled or viable, and empty seeds of three species of Larix tree was studied by transmittance measurements (Tigabu and Oden, 2004). Seeds were classified in those three groups by three PLS-DA models (one for each of the species), with accuracies equal or almost 100% in all independent models. Previously, the same authors had carried out discrimination among empty and viable Pinus Patula seeds, by both transmittance and reflectance (Tigabu and Oden, 2003). Reflectance worked better than transmittance for that application, leading the last one to perfect discrimination by PLS-DA. It was possible to visually discriminate checking the scores of the first PC, possibly due to lipid absorption of filled seeds. Similarly, insect-damaged *Juniperus procera* seeds could be discriminated with a single principal component at 90% accuracy (Tigabu et al., 2007). An application similar to variety discrimination, NIRS with combination of VIS

allowed the discrimination of Pinus Sylvestris seeds by their origin sources and parents with PLS-DA (Tigabu et al., 2005). The accuracies when taking in account the four maternal origin ranged from 80 to 90% (11 PCs); when taking in account parental origin, from 70 to 100% (10PCs). The results reassured that maternal parenting highly influences on the total seed mass. The classification accuracy according to their origin sources was 100% of correctly classified seeds.

JUSTIFICATION FOR WORK AND OBJECTIVES

This dissertation involves the application of NIRS in discrimination of single seeds for quality and safety purposes. Current legislations in countries such as the ones in the European Union have strict standards regarding the presence of genetically modified organisms (GMOs). Current analytical methods for GMO detection are slow and destructive, thus only a small portion of each shipment can be analyzed. The first application in this dissertation, which is the major part, involves the discrimination of Roundup Ready^{TM} single soybeans from conventional using NIR technologies. The discrimination of bulk samples has been previously proved, but not at single seed level. The feasibility of NIR for this purpose could suppose a solution for analyzing entire seed batches and eliminate the sampling effect, resulting in a percentage of GM contamination closer to the actual. On the other hand, a comparison of diverse technologies and classification algorithms in this application is carried out in order to identify the best instrumentation for this purpose.

A couple of applications for quality control involve discrimination of damaged corn kernels. Damaged kernels impact the US grading system and thus decrease the quality of an entire lot. Replacing the current human inspection by a non-destructive automatized method would save time and improve the value of entire seed lots. Because damage may impact seed viability, the feasibility of discrimination viable and non-viable corn kernels and soybean seeds with NIRS is analyzed. Furthermore, the feasibility of this application could be a great tool for seed banks. Seeds kept in storage, even at controlled conditions, age and eventually die. Germoplasms have to closely control the viability of their stored

seeds and replace them once this drops bellows some limits. No practical method exists for breeders to know the percentage of viable seeds in the germoplasm but carrying out periodical accessions and germination tests. This supposes a waste of resources and time. Finally, as the result of the growth of NIR in diverse field and the interest of scientist to learn about its use for future applications, this dissertation provides an insight of the limitations and problems of learning about NIR spectroscopy. Based on the experience acquired in the Grain Quality Laboratory and my earned teaching certificate, a reflection regarding teaching the technologies to young students and the need of suitable material, activities, and guidance is exposed.

The general objectives of this dissertation can be summarized in:

1. Analyze the feasibility, best technologies and algorithms to discriminate Roundup Ready genetically modified soybean seeds from conventional with near infrared spectroscopy
2. Analyze the feasibility of near infrared reflectance spectroscopy to discriminate damaged corn kernels, and viability of corn kernels and soybean seeds.
3. Analyze the current training system in the grain quality laboratory, identify critical points, and update the training material

REFERENCES

Abe, H, Kusama, T., Kawano, S., and Iwamoto, M., 1996. Non-destructive determination of protein content in a single kernel of wheat and soybean by near-infrared spectroscopy. In: Davies, A., Williams, P. (Eds.), The Future waves. NIR publications, Chichester, Wrest Sussex, UK.

Abe, S, and Onishi, K., 2007. Sparse least squares support vector regressors trained in the reduced empirical feature space. Proceedings of 17th International Conference on Artificial Neural Networks ICANN, Porto, Portugal .

Alfassi, Z. B., Boger, Z., and Ronen, Y., 2005. Statistical treatment of analytical data. CRC press, Boca Raton, Florida.

Armstrong, P. R., 2006. Rapid single-kernel NIR measurement of grain and oil-seed attributes. Applied Engineering in Agriculture 22(5), 767-772.

Axun technologies, 2005. Designing a miniature spectrometer, Technical note. Retrieved January 2010 from:
www.axsun.com/.../05-02-084a%20_designing_miniature_spectrometer%20Final.pdf

Bacci, M., Bellucci, C., Cucci, C., Frosinini, C., Picollo, M., Porcinai, S., and Radicati, B., 2005. Fiber optics reflectance spectroscopy in the entire VIS-IR Range: A powerful tool for the non-invasive characterization of paintings. In: Vandiver, P. B., Mass, J. L., Murray, A. (Eds.), Materials Issues in Art and Archaeology VII (vol. 852). Materials Research Society ,Warrendale, PA.

Baianu, I. C., You, T., Costescu, D. M., Lozano, P. R., Prisecaru, V., and Nelson, R. L., 2004. High-resolution nuclear magnetic resonance and near-infrared determination of soybean oil, protein, and amino acid residues in soybean seeds. In: Luthria, D. L. (Ed.), Oil Extraction and Analysis, critical issues and comparative studies. AOCS press, Champaign, IL.

Baker, J. E., Dowell, F. E., and Throne, J. E. , 1999. Detection of parasitized rice weevils in wheat kernels with near-infrared spectroscopy. Biological Control 16, 88-90.

Balas, C., 2009. Review of biomedical optical imaging – a powerful, non-invasive, non-ionizing technology for improving in-vivo diagnosis. Measurement Science Technology 20, 1-12.

Berntsson, O., Danielsson, L-G. , and Folestad, S. , 1998. Estimation of effective sample size when analysing powders with diffuse reflectance near-infrared spectrometry. Analytica Chemica Acta 364(1-3), 243-251.

Brimmer, P. J., DeThomas, F. A., and Hall, J. W., 2001. Method development and implementation of near-infrared spectroscopy in Industrial manufacturing processes. In: Williams, P. C., Norris, K. (Eds.) Near-infrared technology in the agricultural and food industries. AACC press, St. Paul, Minnesota.

Bokobza, L., 1998. Near infrared spectroscopy. Journal of Near Infrared Spectroscopy 6, 3-17.

Börjesson, T., Stenberg, B. , and Schnürer, J. , 2007. Near-Infrared spectroscopy for estimation of ergosterol content in barley: A comparison between reflectance and transmittance techniques. Cereal Chemistry 84(3), 231-236.

Buchanan, R. R., Honigs, D. E., Lee, C. J., and Roth, W., 1988. Detection of ethanol in wines using optical-fiber measurements and near-infrared analysis. Applied Spectroscopy 42, 1106-1111.

Burges, C. J. C., 1998. A Tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121–167

Burns, D. A., and Ciurczak, E. W., 2007. Handbook of near-infrared analysis. Burns, D. A., Ciurczak, E. W. (Eds.). CRC Press, Taylor and Francis group, Boca Raton, Florida.

Campbell, M. R., Sykes, J., and Glover, D. V., 2000. Classification of single- and double-mutant corn endosperm genotypes by near-infrared transmittance spectroscopy. Cereal Chemistry 77(6), 774-778.

Chalmers, J. M., and Dent, G., 2006. Vibrational spectroscopic methods in pharmaceutical solid-state characterization. In: Hilfiker, R. (Ed.), Polymorphism in the pharmaceutical industry. Wiley-VCH, Odense, Denmark.

Cherkassky, V., and Ma, Y., 2002. Practical selection of SVM parameters and noise estimation for SVM regression. University of Minnesota, Minneapolis, Minnesota.

Choquette, S. J., Travis, J. C., Changjiang, Z., & Duewer, D. L., 2002. Wavenumber Standards for Near-infrared Spectrometry. In: Chalmers, J. M. , Griffiths, P. R. (Eds.), Reproduction from Handbook of Vibrational Spectroscopy, John Wiley & Sons Ltd, Chichester, UK. Retrieved 1 September, 2006, from http://www.cstl.nist.gov/acd/839.04/papers/0703_o.pdf#search=%22Wavenumber%20Standards%20for%20Near-infrared%20Spectrometry%22

Chung, H., Ku, M., and Lee, J.,1999. Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene. Vibrational Spectroscopy 22(2), 155-163.

Coats, D. B., 2002. Is near infrared spectroscopy only as good as the laboratory reference values? An empirical approach. Spectroscopy Europe 14(4), 24-26.

Cogdill, R. P., Hurburgh Jr., C. R., and Rippke, G. R., 2004. Single-kernel maize analysis by near-infrared hyperspectral imaging. Transactions of the American Society of Agriculture and Biosystems Engineers 47(1), 311-320.

Cogdill, R. P., Anderson, C. A., Delgado-Lopez, M., Molseed, D., Bolton, R., Herkert, T. , Afnan, A. M. , and Drennen III, J. K. , 2007. Process analytical technology case study part I: Feasibility studies for quantitative near-infrared method development. AAPS: Pharmaceutical Science and Technology 6(2), E262-E272.

Colliez, J., Dufrenois, F., and Hamad, D., 2006. Robust regression and outlier detection with SVR: Application to optic flow estimation. Paper presented to the British Machine Vision Conference (BMVC) , 4-7 September 2006, Edinburgh, UK.

Cortat, F. P. A., 2003. The Kubelka-Munk theory, applications and modifications. Unpublished work, retrieved November 2010 from: http://webstaff.itn.liu.se/~freco/Publications/Courses/Paper_optics_presentation.pdf

Corti, P., Ceramelli, G. , Dreassi, E. , and Mattiim, S. , 1999. Near infrared transmittance analysis for the assay of solid pharmaceutical dosage forms. The Analyst 124, 755-758.

Davies, A.M.C.,and Grant, A., 1987. Review: near infrared analysis of food. International Journal of Food Science Technology 22, 191-207.

Davies, A. M. C., 2005. An introduction to near infrared spectroscopy. Nir News 16(7), 9-21.

Davies, A. M. C., and Fearn, T., 2006. Back to basics: calibration statistics. Spectroscopy Europe, 18(2), 31-32.

Delwiche, S. R., 1993. Measurement of single-kernel wheat hardness using near-infrared transmittance. Transactions of American Society of Agricultural Engineers 36(5), 1431-1437.

Delwiche, S. R., 1995. Single wheat kernel analysis by near-infrared transmittance: protein content. Cereal Chemistry, 72:11-16.

Delwiche, S. R., and Massie, D.R., 1996. Classification of wheat by visible and near-infrared reflectance from single kernels. Cereal Chemistry 73(3), 399-405.

Delwiche, S. R., 1998. Protein content of single kernels of wheat by near infrared reflectance spectroscopy. Journal of Cereal Science, 72:241-254.

Delwiche, S. R. and Hruschka, W. R., 2000. Protein content of bulk wheat from near-infrared reflectance of individual kernels. Cereal Chemistry 77(1), 86-88.

Delwiche, S. R., and Graybosch, R. A., 2002. Identification of waxy wheat by near-infrared reflectance spectroscopy. Journal of Cereal Science 35:29-38.

Delwiche, S. R., Pordesimo, L. O., Scaboo. A. M., and Pantalone, V. R., 2006a. Measurement of inorganic phosphorous in soybeans by near-infrared spectroscopy. Journal of Agriculture and Food Chemistry 54, 6951-6956.

Delwiche, S. R., Graybosch, R. A., Hansen, L. E., Souza, E., and Dowell, F. E., 2006b. Single kernel near-infrared analysis of tetraploid (durum) wheat for classification of the waxy condition. Cereal Chemistry 83(3), 287-292.

Dowell, F. E., 1998. Automated color classification of single wheat kernels using visible and near-infrared reflectance. Cereal Chemistry 75(1), 142-144.

Dowell, F. E., Ram, M. S., and Seitz, L. M., 1999. Predicting scab, vomitoxin, and ergosterol in single wheat kernels using near-infrared spectroscopy. Cereal Chemistry 76(4), 573–576.

Dowell, F. E., Throne, J. E., Wang, D., and Baker, J. E., 1999. Identifying stored-grain insects using near-infrared spectroscopy. Journal of Economic Entomology 92(1),165-169.

Dowell, F. E., Throne, J. E., Broce, A. B., and Xie, F., 2000. Detection of parasitized insects for biological control applications by using NIR spectroscopy. Paper num 003090 for the American Society of Agricultural Engineers annual international meeting, July 9-12, Milwakee, Wisconsin.

Dowell, F. E., Pearson, T. C., Maghirang, E. B., Xie, F., and Wicklow, D. T., 2002. Reflectance and transmittance spectroscopy applied to detecting fumonisin in single corn kernels infected with *Fusarium verticillioides*. Cereal Chemistry 92(2), 222-226.

Dowell, F. E., and Maghirang, E. B., 2002.Novel raw materials, technologies, and products – New challenge for quality control. Paper for presentation at the ICC Conference, May 2002, Budapest, Hungary.

Dowell, F. E., Parker, A. G., Benedict, M. Q., Robinson, A. S., Broce, A. B., and Wirtz, R. A. , 2005. Sex separation of tsetse fly pupae using near-infrared spectroscopy. Bulletin of Entomological Research 95, 249-257.

Dowell, F. E., Maghirang, E. B., Graybosch, R. A., Baenzinger, P. S., Baltensperger, D. D., and Hansen, L. E., 2006. An automated near-infrared system for selecting individual kernels based on specific quality characteristics. Cereal Chemistry 83(5), 537-543.

Despagne, F., and Massart, L., 1998. Neural networks in multivariate calibration. Analyst 123, 157R-178R.

Domanchin, J. , and Gilchrist, J. R. , 2001. Size and spectrum. Photonics July 2001,12-118.

Drucker, H., Burges, C. J.C., Kaufman, L. , Smola, A. , and Vapnik, V. , 1997. Support vector regression machines. Advances in Neural Information Processing Systems 9 , 155-161.

Dryden, G. Mc. , 2003. Near Infrared reflectance spectroscopy: Applications in deer nutrition. Publication No W03/007, Project No UQ 109A. Retrieved February 2009 from http://www.rirdc.gov.au/reports/DEE/w03-007.pdf

Durbha, S. S., King, R., and Younan, N. H., 2007. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Remote Sensing of Environment special issue: Multi-angle imaging spectroradiomenter (MISR) 107(1), 348-361.

Engster, D., and Parlitz, U., 2006. Local and cluster weighted modeling for time series prediction. In: Schelter, B., Winterhalder, M., Timmer, J. (Eds.), Handbook of time series analysis: recent theoretical developments and applications. Wiley-VCH, Germany.

Fearn, T., 2002. Assessing calibrations: SEP, RPD, RER, and R2. NIR News 13(6), 12-14.

Fearn, T., 2005. Chemometrics: an enabling tool for NIR, NIR news 16(7), 17-19.

Finney, E. E., and Norris, K. H., 1978. Determination of moisture in corn kernels by near-infrared transmittance measurements. Transactions of the American Society of Agricultural Engineers 21, 581-584.

Frank, I. E., and Lanteri, S., 2001. Classification models: discriminant analysis, SIMCA, CART. Chemometrics and Intelligent Laboratory Systems 5(3), 247-256.

Frohlich, H., and Zell, A., 2005. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. Proceedings of International Joint Conference on Neural Networks, Montreal, Canada.

Garini, Y., Young, I. T., and McNamara, G. , 2006. Spectral imaging: Principles and applications, Cytometry 69A(8), 735 – 747.

Geladi, P., Burger, J., and Lestander, T., 2004. Hyperspectral imaging: calibration problems and solutions. Chemometrics and Intelligent Laboratory Systems 72, 209-217.

Ghaedian, A. R., and Wehling, R. L., 1997. Discrimination of sound and granary-weevil-larva-infested wheat kernels by near-infrared diffuse reflectance spectroscopy. Journal of American Oil Association of Chemists International 80 (5), 997–1005.

Greensill, C. V., and Walsh, K. B. , 2000. Optimization of instrument precision and wavelength resolution for the performance of NIR calibrations of sucrose in water –cellulose matrix. Applied Spectroscopy, 54(3), 426-430.

Herschel, F. W., 1800. Investigation of the powers of the prismatic colours to heat and illuminate objects. Philosophical Transactions of the Royal Society of London 90, 255-329.

Haaland , D. M. , and Thomas, E. V. , 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Analytical Chemistry 60, 1193-1202.

Hacisalihoglu, G., Larbi, B., and Settles, A. M., 2010. Near-infrared reflectance spectroscopy predicts protein, starch, and seed weight in intact seeds of common

bean (phaseolus vulgaris, L.). Journal of Agriculture and Food Chemistry 58, 702-706.

Hammateenejad, B., Akhind, M., and Samar, F., 2007. A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: Effect of wavelength selection. Spectrochimica Acta, part A: Molecular and Biomolecular Spectroscopy 67(3-4), 958-965.

Herschel, F. W., 1800. Investigation of the powers of the prismatic colours to heat and illuminate objects. Philosophical Transactions of the Royal Society of London 90, 255-329.

Holler, S. , Pan, Y. , Chang, R. K. , Bottiger, J. R. , Hill, S. C. , and Hillis, D. B. , 1998. Two-Dimensional Angular Optical Scattering for the Characterization of Airborne Microparticles. Optics Letters 23, 1489–1491.

Hom, N. H., Becker, H. C., and Mollers, C., 2007. Non-destructive analysis of rapeseed quality by nirs of small seed samples and single seeds. Euphytica 153, 27-34.

Hunt, W. H. , Fulk, D. W., Elder, B. and Norris, K., 1997. Collaborative study on near infrared reflectance devices for determination of protein in hard red winter wheat, and for protein and oil in soybeans. Cereal Foods World, 22:534-536, 538.

Hurburgh Jr, C. R., 1989. Agricultural engineering staff papers series FPR 89-2. Amearican Society of Agriculture Engineers meeting presentation, St. Joseph, June 25-28 1989, 1-13.

Hopkins, D. W., 2008. Using data pretreatments effectively. Seminar at International Diffuse Reflectance Conference (IDRC), Chamberburgh ,Pennsylvania.

Howard, M., and Workman Jr., J., 2005. Chemometrics in spectroscopy-Linearity in calibration: The Durbin-Watson statistic, Spectroscopy 20(3), 34-40.

Hruschka, W. R., 1987. Data analysis: Wavelength selection methods. In Williams, P., Norris, K. (Eds.), Near–Infrared Technology in Agricultural and Food Industries , American Association of Cereal Chemists , St. Paul, Minnessota

Hsu, C., Chang, C. and Lin, C., 2003. A practical guide to support vector classification. Retrieved January 2010 from: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Hubert, M. , Rousseeuw, P. J. , and Van Aelst, S. , 2008. High-breakdown robust multivariate methods. Statistical Science 23(1), 92-119.

Ingle, J. D. Jr., and Crouch, S. R. , 1988. Spectrochemical Analysis, Prentice Hall, Upper saddle river, New Jersey.

Janni, J., 2007. Pionerr Hi-Bred International Inc. Patent 7274457 B2

Janni, J., Weinstock, B. A., Hagen, L., and Wright, S., 2008. Novel near-infrared sampling apparatus for single kernel analysis of oil content in maize. Applied Spectroscopy 62(4), 423-426.

Jiao, Z., Si, X., Li, G., Zhang, Z., and Xu, X., 2010. Unintended compositional changes in transgenic rice seeds (Oryza sativa L.) studied by spectral and chromatographic analysis coupled with chemometric methods. Agricultural and Food Chemistry 58, 1746-1754.

Kalivas, J. H., and Gemperline, P. J. , 2006. Calibration. In: Gemperline, P. J. (Ed.), Practical guide to chemometrics. CRC Taylor and Francis group, Boca Raton, Florida.

Kavraki, L. E., 2007. Dimensionality Reduction Methods for Molecular Motion. Module in Connexions website. Retrieved November 2010 from: http://cnx.org/content/m11461/latest/

Kays, S. E. , Shimizu, N. , Barton II , F. E. , and Ohtsubo, K. , 2005. Near-Infrared transmission and reflectance spectroscopy for the determination of dietary fiber in barley cultivars. Crop Science 45, 2307-2311.

Kovalenko, I. V. , Rippke, G. R. , and Hurburgh, C. R. , 2006. Determination of amino acid composition of soybeans (Glycine max) by near-infrared spectroscopy. Journal of Agricultural and Food Chemistry 54, 3485–3491.

Kusama, T. , Abe, H., Kawano, S., Iwamoto, M., 1997. Classification of normal and aged soybean seeds by Discriminant Analysis using principal component scores of near infrared spectra. Nippon Shokuhin Kogyo Gakkai-Shi 44(8), 569-578.

Lamb, D. T. and Hurburgh, C. R., 1991. Moisture determination in single soybean seeds by near-infrared transmittance. Transactions of the American Society of Agricultural Engineers, 34(5), 2123-2129.

Lavine, B. K., 2000. Clustering and classification of analytical data. In: Meyers, R. A. (Eds.), Encyclopedia of analytical chemistry. Wiley & Sons Ltd., Chichester, UK.

Maghirang, E. B., and Dowell, F. E., 2003. Hardness measurement of bulk wheat by single-kernel visible and near-infrared reflectance spectroscopy. Cereal Science 80(3), 316-322.

Maghirang, E. B., Dowell, F. E., Baker, J. E., and Throne, J. E., 2003. Automated detection of single wheat kernels containing live or dead insects using near-infrared reflectance spectroscopy. Transactions of the American Society of Agricultural Engineers 46(4), 1277-1282.

Mahesh, S., Manickavasagan, A., Jayas, D. S., Paliwal, J., and White, N. D. G., 2008. Feasibility of near-infrared hyperspectral imaging to differentiate Canadian wheat classes. Biosystems Engineering 101, 50-57.

Malinen, J. , Känsäkoski, M. , Rikola, R. , and Eddison, C. G. , 1998. LED-based NIR spectrometer module for hand-held and process analyzer applications. Sensors and Actuators B: Chemical 51(1-3), 220-226.

Manley, M., Williams, P., Nilsson, D., and Geladi, P., 2009. Near infrared hyperspectral imaging for the evaluation of endosperm texture in whole yellow maize (zea maize L.) kernels. Journal of Agriculture and Food Chemistry 57, 8761-8769.

Massart, D. L., Vandegiste, B. G. M., Buydens, L. M. C., De Jong, S., Lewi, P. J., and Smeyers-Verbeke, J., 1998. Handbook of Chemometrics and Qualimetrics, Part B. Elsevier, Amsterdam, Netherlands.

McClure, W. F. , Moody, D. , Standfield, D. L. , and Kinoshita, O. , 2002. Hand-held NIR spectrometry. Part II: An economical no-moving parts spectrometer for measuring chlorophyll and moisture. Applied Spectroscopy 56(6), 720-724.

Momma, M., and Bennett, K., 2002. A pattern search method for model selection of support vector regression. In Proceedings of the SIAM International Conference on Data Mining, Philadelphia, Pennsylvania.

Muñiz, R., Perez, M. A. , De la Torre, C. , Carleos, C. E. , Corral, N. , and Baro, J. A. , 2009. Comparison of principal component regression (PCR) and partial least square (PLS) methods in prediction of raw milk composition by VIS-NIR

spectrometry. Application to development of on-line sensors for fat, protein and lactose contents. Oral presentation proceedings for XIX IMEKO world congress of applied Metrology, Lisbon, Portugal. Retrieved January 2010 from http://www.imeko2009.it.pt/Papers/FP_229.pdf

Naes, T., Irgens, C., and Martens, H., 1986. Comparison of linear statistical methods for calibration of NIR instruments. Journal of the Royal Statistical Society Series C (Applied Statistics) 35(2), 195-206.

Naes, T., 1987. The design of calibration in near reflectance analysis by clustering, Journal of Chemometrics 1, 121-134.

Naes, T., 1991. Multivariate calibration when data are split into subsets. Journal of Chemometrics 5, 487-501.

Naes, T. , Isaksson, T., Fearn, T., and Davies, T., 2002a. Selection of samples for calibration. In: A User-Friendly Guide to Multivariate Calibration and Classification. NIR publications, Chichester, UK.

Naes, T., Isaksson, T., Fearn, T., and Davies, T., 2002b. Outlier detection. In: A User-Friendly Guide to Multivariate Calibration and Classification. NIR publications, Chichester, UK.

Naes, T.,Isaksson, T., Fearn, T., and Davies, T., 2002c. Multiplicative scatter correction (MSC). In: A User-Friendly Guide to Multivariate Calibration and Classification, NIR publications. Chichester, UK.

Naes, T., Isaksson, T., Fearn, T., and Davies, T., 2002d. Non-linearity problems in calibration, in A User-Friendly Guide to Multivariate Calibration and Classification. NIR publications, Chichester, UK.

Naes,T., Isaksson, T., Fearn, T., and Davies, T., 2002e. Validation, in: A User-Friendly Guide to Multivariate Calibration and Classification. NIR publications , Chichester, UK.

Nielsen, J. P., Pedersen, D. K., and Munck, L., 2003. Development of nondestructive screening methods for single kernel characterixzation of wheat. Cereal Chemistry 80(3), 274-280.

Niewitetzki, O., Tillman, P., Becker, H. C., and Mollers, C., 2010. A new near-infrared reflectance spectroscopy method for high-throughput analysis of oleic acid and linolenic acid content of single seeds in oilseed rape (brassica napus L.). Journal Agriculture and Food Chemistry 58, 94-100.

Norris, K., and Williams, P. C. , 1984. 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 Chemistry 61, 158-165.

Orman, B. A., and Schumann Jr, R. A., 1992. Nondestrctive single-kernel oil determination of maize by near-infrared transmission spectroscopy. Journal of the American Oil Chemical Society 69(10),1036-1038.

Pasikatan, M. C., and Dowell, F. E., 2004. High-speed NIR segregation of high and low-protein single wheat seeds. Cereal Chemistry 81(1),145-150.

Pearson, T. C., Wicklow, D. T., Maghirang, E. B., Xie, F., and Dowell, F. E., 2001. Detecting aflatoxin in single corn kernels by transmittance and reflectance spectroscopy. Transactions of the American Society of Agricultural Engineers 44(5), 1247-1254.

Perez-Mendoza, J., Flinn, P. W., Campbell, J. F., Hackstrum, D. W., and Throne, J. E., 2004. Detection of stored-grain insect infestation in wheat transported in railroad hopper-cars. Journal of Economic Entomology 97(4), 1474-1483.

Pou Saboya, N. , 2002. Analisi de control de preparados farmaceuticos mediante espectroscopia en el infrarojo proximo. Unpublished Phd thesis, Universitat Autonoma de Barcelona, Barcelona, Spain.

Prieto, N., Roehe, R., Lavin, P., Baeten, G., and Andres, S. , 2009. Application of near infrared reflectance spectroscopy to predict meat and meat products quality: a review. Meat Science 83(2), 175-186.

Reeves, J. B., and Zapf, C. M. , 1998. Mid-Infrared diffuse reflectance spectroscopy for discriminant analysis of food ingredients. Journal of Agriculture and Food Chemistry 46, 3614-3622.

Richardson, A. D., and Reeves III, J. B., 2005. Quantitative reflectance spectroscopy as an alternative to traditional wet lab analysis of foliar chemistry: near-infrared and

mid-infrared calibrations compared. Canadian Journal for Forest Research 35, 1122-1130.

Ridgway, C., and Chambers, J., 1998. Detection of insects inside wheat kernels by NIR imaging. Journal of Near Infrared Spectroscopy 6(1), 115–119.

Ridgeway, C., and Chambers, J., 1999. Detection of grain weevils inside single wheat kernels by a very near infrared two-wavelength model. Journal of Near Infrared Spectroscopy 7(4), 213–221.

Robertson, J. A., and Windham, W. R., 1981. Comparative study of methods of determining oil content of sunflower seed. Journal of the American Oil Chemists' Society 58(11), 993-996.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmon, A., and Jent, N., 2007. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. Journal of Pharmaceutics and Biomedical Analysis 44(33), 683-700.

Sarraguca, M. C. , Paulo, A. , Alves, M. M. , Dias, A. M. A. , Lopes, J. A. , and Ferreira, E. C. , 2009. Quantitative monitoring of an activated sludge reactor using on-line UV-visible and near-infrared spectroscopy. Analytical and Bioanalytical Chemistry 395(4), 1159-1166.

Sato, T., Takahata, Y., Noda, T., Yanagisawa, T., Morishita, T., and sakai, S., 1995. Nondestructive determination of fatty acid composition of husked sunflower (helianthus annua L.) seeds by near-infrared spectroscopy. Journal of American Oil Chemists Society, 72(10), 1177-1183.

Sato, T., Uezono, I., Morishita, T., and Tetsuka, T., 1998. Nondestructive estimation of fatty acid composition in seeds of brassica naptus L. by near-infrared spectroscopy. Juornal of American Oil ChemistsSociety, 75(12), 1877-1881.

Schumann, A. W. , and Meyer, J. H. , 2000. Progress with the implementation of diode array near-infrared spectrometer for direct at-line analysis of sugarcane samples. Proceedings of South Africa Sugar Technology Association 74, 122- 123.

Short, S. M. , Cogdill, R. P. , and Anderson, C. A. , 2008. Figures of merit comparison of reflectance and transmittance near-infrared methods for the prediction of

constituent concentrations in pharmaceutical compacts. Journal of Pharmaceutical Innovation 3(1), 41-50.

Silvela, L., Rogers, R., Barrera, A., and Alexander, D. E., 1989. Theorethical Applications of Genetics 78, 298.

Singh, C. B., Jayas, D. S., Paliwal, J., and White, N. D. G., 2009. Detection of insect-damaged wheat kernels using near-infrared hyper spectral imaging. Journal of Stored Products Research 45, 151-158.

Singh, C. B., Jayas, D. S., Paliwal, J., and White, N. D. G., 2010. Identification of insect-damaged wheat kernels using short-wave near-infrared hyper spectral and digital colour imaging. Computers and Electronics in Agriculture 73, 118-125.

Smail, V. W., Fritz, A. K., Wetzel, D. L., 2006. Chemical imaging of intact seeds with NITR focal plane array assists plant breeding. Vibrational Spectroscopy 42, 215-221.

Smith, J. P. , 2000. Product review: Spectrometers get small. Miniature spectrometers rival benchtop instruments. Analytical Chemistry, 72(19):653A-658A.

Smith, B.C. , 2002. Fundamentals of molecular absorption spectroscopy. In Quantitative spectroscopy: theory and practice, Academic Press, San Diego, California.

Spielbauer, G., Amstrong, P., Baier, J. W., Allen, W. B., Richradson, K., Shen, B., and Settles, M., 2009. High-throughput near –infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. Cereal Chemistry 86(5), 556-564.

Stark, E. and Luchter, K. , 2005. NIR instrumentation technology, NIR News 16(7), 13-16.

Sternberg, J. C. , Stillo, H. S. , and Schwendeman, R. H. , 1960. Spectrophotometric analysis of multicomponent systems using the least squares method in matrix form. Analytical Chemistry 32, 84-90.

Tajuddin, T., Watanabe, S., Masuda, R., Harada, K., and Kawano, S., 2002. Application of near infrared transmittance spectroscopy to the estimation of protein and lipid contents in singke seeds of soybean recombinant inbred lines for quantitative trait loci analysis. Journal of Near Infrared Spectroscopy 10(4), 315-325.

Tallada, J. G., Palacios-Rojas, N., and Amstrong, P. R., 2009. Prediction of maize seed attributes using a rapid single kernel near infrared instrument. Journal of Cereal Science 50, 381-387.

Tamburini, E. , Vaccari, G. , Tosi, S. , and Trilli, A. , 2003. Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fiber probe. Applied Spectroscopy 57(2), 132-138.

Thermo Fisher Scientific, 2006. Advantages of Fourier-Transform Near-Infrared Spectroscopy, Application note 50771, retrieved January 2010 from: http://www.thermo.com/eThermo/CMA/PDFs/Articles/articlesFile_1195.pdf

Tigabu, M., and Oden, P. C., 2003. Discrimination of viable and empty seeds of pinus patula schiede and deppe with near infrared spectroscopy. New Forests 25, 163-176.

Tigabu, M., and Oden, P. C., 2004. Simultaneous detection of filled, empty and insect-infested seeds of three Larix species with single seed near-infrared transmittance spectroscopy. New Forests 27, 39-53.

Tigabu, M., Oden, P. C., and Lindgren, D., 2005. Identification of seed sources and parents of pinus sylvestris L. using visible-near infrared reflectance spectra and multivariate analysis. Trees 19, 468-476.

Tigabu, M., Fjellstrom, J., Oden, P. C., and Teketay, D. , 2007. Germination of *Junisperus Procera* seeds in response to stratification and smoke treatments, and detection of insect-damaged seeds with VIS+NIR spectroscopy. New Forests 33, 155-169.

Tillman, B. L., Gorbet, D. W., and Person, G., 2006. Predicting oleic and linoleic acid content of single peanut seeds using near-infrared reflectance spectroscopy. Crop Science 46, 2121-2126.

Tran, T. N., Wehrens, R., and Buydens, L. M. C., 2005. Clustering multispectral images: a tutorial. Chemometrics and Intelligent Laboratory Systems 77, 3– 17.

Vapnik, V., and Lerner, A., 1963. Pattern recognition using generalized portrait method. Automation and Remote Control 24, 774–780.

Vapnik, V., 1995. The nature of statistical learning theory. In: Jordan, M., Lawless, J. F., Lauritzen, S. L., Nair, V. (Eds.).Springer Verlag, New York.

Velasco, L., Perez-Bich, B., and Fernandez-Martinez, J. M., 1999a. Use of near-infrared reflectance spectroscopy for selecting for high stearic acid concentration in single husked achenes of sunflower. Crop Science 39, 219-222.

Velasco, L., Mollers, C., and Becker, H. C., 1999b. Estimation of seed weight, oil content and fatty acid composition in intact single seeds of rapeseed (Brassica napus L.) by near infrared reflectance spectroscopy. Euphytica 106, 79-85.

Velasco, L., and Mollers, C., 2002. Nondestructive assessment of protein content in single seeds of rapeseed (Brassica napus L.) by near-infrared reflectance spectroscopy. Euphytica 123, 89-93.

Velasco, L., Perez-Vich, B., and Fernandez-Martinez, J. M., 2004. Use of near-infrared reflectance spectroscopy for selecting for high stearic acid concentration in single husked achenes of sunflower. Crop Science 44, 93-97.

Walczak, B. , and Massart, D. L. , 1998. Multiple outlier detection revisited. Chemometrics and Intelligent Laboratory Systems 41(1), 1-15.

Wang, D., Dowell, F. E., and Chung, D. S., 2001. Assessment of heat-damaged wheat kernels using near-infrared spectroscopy. Presentation at the 2001 American Society of Agriculture Engineers annual international meeting, sacramento, Ca, july 30-august 1.paper num. 01-6006.

Wang, D., Dowell, F. E., and Dempster, R., 2002a. Determining vitreous subclasses of hard red spring wheat using visible/near-infrared spectroscopy. Cereal Chemistry 79(3), 418-422.

Wang, D., Ram, M. S., and Dowell, F. E., 2002b. Classification of damaged soybean seeds using near-infrared spectroscopy. Transactions of the American Society of Agriculture Engineers 45(6), 1943-1948.

Wang, D., Dowell, F. E., Ram, M. S., and Schapaugh, W. T., 2003. Classification of fungal-damaged soybean seeds using near –infrared spectroscopy. International Journal of Food Properties 7(1), 75-82.

Wang, W., and Paliwal, J. , 2006. Design and evaluation of a visible-to-near-infrared electronic slitless spectrograph. Science Technology 17, 2698-2704.

Weinstock, B. A., Janni, J., Hagen, L., and Wright, S., 2006. Prediction of oil and oleic acid concentrations in individual corn (zea mays L.) kernels using near-infrared reflectance hyperspectral imaging and multivariate analysis. Applied Spectroscopy 60(1), 9-16.

Wilkin, D. R., and Fleurat-Lessard, F., 1990. The detection of insects in grain using conventional sampling spears. In: Fleurat-Lessard, F., Ducom, P. (Eds.), Proceedings of the 5th international working conference stored-product protection III, Bordeaux, France.

Williams, P. C., and Sovering, D. C. , 1993. Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. Journal of Near Infrared Spectroscopy 1, 25-32

Williams, P. C., 2001. Implementation of near infrared technology. In: Williams, P. C., Norris, K. (Eds.), Near-Infrared technology in the agricultural and food industries. AACC, St Paul, Minnesota.

Williams, P., Geladi, P., Fox, G., and Manley, M., 2009. Maize kernel hardness classification by near infrared (NIR) hyperspectral imaging and multivariate data analysis. Analytica Chimica Acta 653, 121-130.

Wilson, R. H., and Tapp, S. H., 1999. Mid-infrared spectroscopy for food analysis: recent new applications and relevant developments in sample presentation methods. Trends in Analytical Chemistry 18(2), 85-93.

Windig, W., Shaver, J., and Bro, R., 2008. Loopy msc: a simple way to improve multiplicative scatter correction. Applied Spectroscopy 62(10), 1153-1159.

Wise, B., no date. Properties of partial least squares regression and differences between algorithm. Brochure retrieved January 2010 from: http://www.eigenvector.com/Docs/Wise_pls_properties.pdf

Wold, S., 1975. Soft modeling by latent variables; the non-linear iterative partial least squares approach. In: Gani, J. (Ed.), Perspectives in probability and statistics, papers in honour of M.S. barlett, Academic Press, London, UK.

Wold, S., 1976. Pattern recognition by means of disjoint principal components models. Pattern Recognition 8, 127–139.

Wold, S., Eriksoon, L., Trygg, J., and Kettaneh, N., 2004. The PLS method - partial least squares projections to latent structures- and its applications in industrial RDP (research, development, and production). Retrieved January 2010 from: http://www.umetrics.com/pdfs/events/prague%200408%20__%20PLS_text_wold.pdf

Workman, J. J. , 2005. An introduction to Near Infrared spectroscopy. Retrieved January 2010 from: http://www.spectroscopynow.com/coi/cda/detail.cda?id=1881&type=EducationFeature&chId=2&page=1

Wright, S., 2007. Pioner Hi-Bred International Inc. Patent 7274456 B2

Yu, B. , Burnside, E. S. , Sisney, G. A. , Harter, J. M. , Changfang, Z. , Dhalla, A., and Ramunjam, N. , 2007. Feasibility of near-infrared diffuse optical spectroscopy on patients undergoing image guided core-needle biopsy. Optics Express 15(12), 7335-7350.

Zomer, S., Brereton, R.G., Carter, J.F., and Eckers, C., 2004. Support vector machines for the discrimination of analytical chemical data: application to the determination of tablet production by pyrolysis-gas chromatography-mass spectrometry. Analyst 129.

CHAPTER 2. NEAR INFRARED REFLECTANCE SPECTROSCOPY APPLIED TO DISCRIMINATION OF CONVENTIONAL AND ROUNDUP READY SOYBEAN SEEDS

A paper to be submitted to the Journal of Applied Spectroscopy

[1]Lidia Esteve Agelet, [2]Aoife A. Gowen, [1]Charles R. Hurburgh*, [2]Colm P. O'Donell

[1]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011

[2] Agricultural and Food Science Centre, University College of Dublin, Belfield, Dublin 4, Republic of Ireland

* To whom correspondence should be addressed at 1547 Food Science Building, Iowa State University, Ames, IA 50011-1060. Phone: 515-294-8629, Fax: 515-294-6383, Email: tatry@iastate.edu

ABSTRACT

Identification and proper labeling of genetically modified organisms is required and increasingly demanded by law and consumers worldwide. In this study, the feasibility of near infrared reflectance technologies for discriminating Roundup Ready® and conventional (not genetically modified) soybean (Glycine max L.) seeds is studied. Over 200 seeds of each class (Roundup Ready® and conventional) were used. A low resolution pushbroom imaging unit, a commercial diode array instrument with single seed adapter, and a non-commercial instrument (light tube) which takes the whole seed reflectance spectra were tested. Principal Component Analysis with Artificial Neural Networks (PCA-ANN) and Locally Weighted Principal Component Regression (LWR-PCR) were used for creating the discrimination models. Imaging unit classification accuracies around

89% were achieved with both PCA-ANN and LW-PCR when validation was performed with seeds belonging to samples and images included in the training set. Independent validation with new images with a percentage of seeds from samples not included in the training set gave accuracies around 75%. PCA-ANN models for both single point instruments lead to accuracies in the low 80 percent range when validated with seeds belonging to samples included in the training set. LWR-PCR models had higher accuracies (over 90%). The light tube outperformed the two other instruments due to its sensitivity to seed size and shape. The overall accuracies dropped below 80% when new models were validated with seeds from samples not represented in the training set. These results show the ability of NIRS reflectance technologies to discriminate individual Roundup Ready and conventional seeds at good screening accuracies whenever the seeds belong to a sample already included in the training set.

KEYWORKDS: Roundup Ready soybeans, Near Infrared, discrimination, imaging, reflectance.

INTRODUCTION

In 1994, the first genetically modified (GM) soybean (*Glycine max L.)* variety was introduced in the United States market. They were the first generation of Roundup Ready® soybeans that incorporates a gene from the bacterium *Agrobacterium Tumefaciens* conferring resistance to the glyphosate Roundup-brand herbicide. Although other GM soybean varieties are expected in the future– with genes conferring high oleic acid or resistance to other herbicides - the most widespread GM soybean varieties at present are Roundup Ready ®, comprising close to 60% of worldwide soybean crops.[1] Despite rapid acceptance controversy remains. The uncertainty of GM effects on human health, environmental safety, and ecological quality (i.e. varietal preservation) are some of the concerns associated with GM technologies.

Low-level mixtures of GM in conventional batches of soybeans, for instance by incomplete cleaning of machinery during harvesting operations, are common and

virtually impossible to eliminate.[2] With the increasing demand for organic products, worldwide governments have created regulations for labeling and control of GM products. Worldwide acceptance thresholds of adventitious GM contamination in soybeans range from 1 to 5%. Below those thresholds, there is no need for specific labeling of soybean batches destined to both human and animal feeding purposes, as there is no rule for GM-fed animal products for human consumption.[3] The European Union has the smallest tolerance for GM admixture. The European Novel Food Regulation EC 1829/2003 sets a threshold of 0.9 % of GM contamination in food and animal feed without labeling if the adventitious contamination comes from one of the accepted GM varieties (2 soybean varieties so far, Roundup Ready® being one of them). A 0.5% tolerance limit is applied for unauthorized GM varieties if they are proved safe by relevant scientific committees.

The need for GM detection methods that are accurate, fast, and inexpensive remains. Currently, there are two recognized classes of GM identification methods, both based on detecting two molecules: DNA and protein. Polymerase Chain Reaction (PCR) methods are the most sensitive, with lower limits of GM DNA detection of 0.1%.[4] Protein-based detection methods such as Enzyme-Linked Immunosorbent Assay (ELISA) are less accurate; more sample is needed, and prior knowledge of the GM protein (or specific GM event) is required. However, protein-based methods are faster and simpler to perform. Other methods are seed germination, tetrazolium tests, insect resistance bioassays, biosensors, chromatography, use of microfabricated devices, and nanoscale analysis.[5] All require sample destruction, considerable time, and human resource expenses. Because only few seeds can be taken per analysis, the accuracy of the method becomes dependant on the sampling procedure. The distribution of GM seeds in a conventional batch of any grain can create up to 20% sampling error.[6]

Roussel et al.[7] introduced the use of Near Infrared Spectroscopy for discrimination of conventional and Roundup Ready® bulk soybeans by Near Infrared Spectroscopy (NIRS) transmittance. Three classification methods (Partial Least Squares Discrimination Analysis (PLS-DA), Artificial Neural Networks (ANN), and Locally Weighted Regression (LWR)) were tested. Non-linear methods, ANN and LWR, achieved the

highest classification accuracies of 88 and 93% respectively. Near Infrared measurements do not measure compounds at trace levels such as part per million, but could measure the physical or chemical expression of a genetic trait. In this case, it was suspected that differences among Roundup Ready® and conventional soybeans arose from fiber structure, after observing the relevance of the regression coefficients from the PLS-DA models in the carbohydrate absorbance region (894 - 950 nm).

Near Infrared technologies are known to provide fast and non-destructive analysis, which offers an attractive way to measure whole batches of seeds and reduce sampling limitations. The previous study[7] was carried out on bulk samples, which implies that a whole sample of around 250 – 500 grams of seeds was scanned and classified as either Roundup Ready® or conventional. The threshold or percentage of sample impurity was unknown. In this study, we use Near Infrared Reflectance Imaging (NIR-CI) and two single point NIRS reflectance instruments to discriminate GM and conventional single soybean seeds. Since reflectance measurements are especially adequate for online measurements, the success of this application would provide a fast and inexpensive method to detect Roundup Ready® impurities expressed as proportion or number of GM seeds in whole batches of conventional soybeans. There were two main objectives in this study: 1) to study the feasibility of NIRS reflectance technologies for discrimination of Roundup-Ready® and conventional soybean single soybean at the single seed level. 2) to compare possible differences in classification accuracies from technologies (imaging versus single point instruments) and instrumentation or sampling method (whole seed versus half seed).

EXPERIMENTAL

Chemical Imaging Unit

Instrumentation. The reflectance imaging system used for this study was a line-scanning instrument (DV Optics Ltd., Padua, Italy) with an InGaS camera with low resolution (320 x 240 pixels, 12 bits resolution) and a Specim N17E spectrograph (Spectral Imaging Ltd., Oulu, Finland). The wavelength range covered was from 880 nm to 1,720 nm, taking data

points every 7 nm. The translation stage located under the camera where seeds were placed was set to a speed equal to 20 µm/s, obtaining images of 350 lines by 320 columns. The seeds were arranged in batches of 60 seeds on the instrument translation stage (FIG. 1).

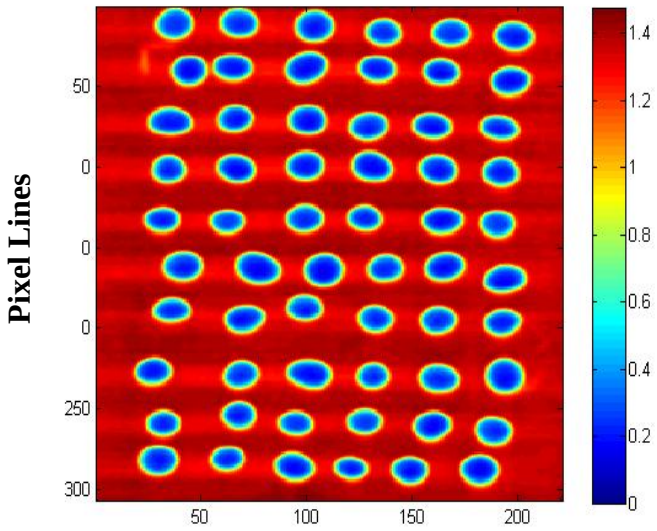


FIG.1. Absorbance image showing the placement of 60 soybean seeds. The image has been reduced to 310 long (lines) x 220 pixels wide (columns) to remove empty pixels.

Samples and scanning. We picked 216 Roundup Ready® (RR) and 202 conventional soybean samples from the Grain Quality Laboratory storage bank at Iowa State University (Ames, IA), covering crop years from 1984 to 2008. Fifteen individual seeds were randomly taken from each sample and mixed, obtaining two bags of 3,240 conventional and 3,030 RR seeds. A first set of 92 images were obtained scanning 30 RR and 30 conventional seeds per image, drawn from the two bags (RR and conventional) and randomly arranged on the instrument translation stage as shown in figure 1.

We collected a second set of 31 images with a small and variable proportion of Roundup Ready® seeds, simulating a real situation of screening for Roundup Ready®

contamination of conventional batches (TABLE I). Again, a total of 60 seeds were arranged per image. Twenty new RR samples, not used in the previous 92 images, were selected. Thirteen seeds were drawn from each RR sample and mixed in a bag. The conventional seeds used in those 31 images were taken from a mixture of 45 previously used samples (20 new seeds drawn per sample) plus a set of 25 new conventional samples (22 seeds/sample). Summarizing, all RR seeds and approximately ¼ of the conventional seeds in the 31 images belonged to new samples not included in the initial 92 images.

TABLE I. Composition of the 31 images used as validation set with the number of seeds from each category (conventional and RR) per image and number of images

Number of 60-seed images	**6**	**6**	**6**	**7**	**6**
Number of Conventional seeds	59	58	55	50	40
Number of Roundup Ready® seeds	1	2	5	10	20

Image and Spectra Processing. Matlab v.7.4 (Mathworks, Natick, MA) and PLS_toolbox v.3.5.4 (Eigenvector Research, Inc., Wentachee, WA) were used for collecting and preprocessing the images. Individual seed spectra were retrieved by building an absorbance mask at 1,048 nm; the absorbance threshold was selected after series of small tests to achieve a conservative value which excluded the seed edges with high light scattering and background. Each seed was represented by approximately 150 pixels (150 spectra) on average. The spectra belonging to each pixel from a seed was preprocessed by multiplicative scatter correction (MSC)[8] using the individual seed mean absorbance spectrum (average of all pixel spectra for that seed) as reference. The overall average from the MSC preprocessed spectra was taken to obtain one spectrum per seed. This method has enhanced the signal and reduced curvature effect in spherical objects. [9] The

working wavelength range was reduced to the region of 943 - 1,643 nm to reduce noise. Noise peaks at 1,321 nm were observed for some of the spectra, possibly due to light scattering effects on instrument conformation. Principal component analysis (PCA) was carried out on the spectra from 1300 nm to 1405 nm. Spectra with the noise peak characteristic at 1,321 nm showed high score values on the third principal component (FIG.2), allowing efficient identification and removal of those spectra. The final data set from the initial 92 images had 5,222 spectra (5% of the total data removed as potential outliers). From the next 31 images, the final data set had 1765 spectra (5% of data removed as potential outliers, again).

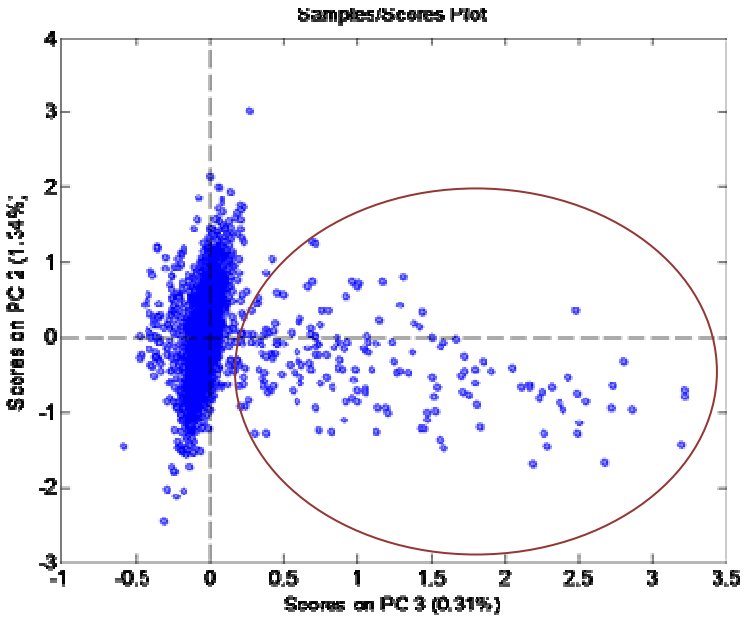


FIG.2. Sample scores on second (PC2) and third (PC3) principal components from PCA carried on single seed spectra from 1300 nm to 1405 nm. Circled samples showing high scores on PC3 belonged to noisy spectra that were removed.

Mathematics. Following the previous work at Roussel et al.[7], we used non-linear classification methods for this study: Locally Weighted Principal Component Regression (LW-PCR), and Principal Component Analysis followed by Artificial Neural Networks (PCA-ANN). We utilized Matlab v.7.04 (Mathworks, Natick, MA) combined with the

PLS_toolbox v.3.5.4 (Eigenvector Research Inc., Wenatchee, WA) functions for data analysis and model development. LW-PCR models were based on the original algorithm provided by PLS toolbox v.2.1.1 (Eigenvector Research Inc., Wenatchee, WA) in 2000. In this algorithm, each new sample category is determined according to a reduced optimized number of neighbors or closest samples according to spectral similarity. The closeness between spectra is defined by Mahalanobis distance in the first principal component and their relevance in the classification of the new sample is defined by a cubic weight function. The closest neighbors have more relevance when determining the new sample category by Principal Component Regression (PCR) in the reduced neighborhood. The number of principal components for PCR and neighbors were optimized by iteration, across a range from 20 to 1000 neighbors at increments of 20, and from 8 - 20 PCs.

Artificial Neural Networks (ANN) simulates the human nervous system regarding data management and learning procedure. ANN models were created and trained by feed-forward backpropagation, meaning that according to the error calculated after each training session (epoch) the weights of each neuron are adjusted. The best training function was chosen among those offered by the ANN toolbox for Matlab v.7.04 (Mathworks, Natick, MA): resilient backpropagation algorithm, Bayesian regulation backpropagation, and Levenberg-Marquardt backpropagation. The number of input neurons, as with the number of PCs, was optimized by iteration from 10 to 20. The number of neurons in the hidden layer was also optimized, testing from 2 to 4. Transfer functions connecting the neurons between the input and hidden layer were hyperbolic tangent sigmoids. The connection between hidden and output layer, with two neurons corresponding to the two classification categories, was done by linear transfer functions. The seed classes were coded (1,0) for conventional and (0,1) for RR seeds. The number of learning epochs were monitored and optimized to avoid overfitting. The weights were reset to zero ten times for each combination of parameters to avoid local minima, and best final model was selected according to the lowest number of misclassified samples from the validation set.

Discrimination Models. Two types of models were created in order to test the generalization ability of the algorithms to new seeds coming from new samples and new images. For the first models, data belonging to the first set of 92 images were taken for both training and validation. For PCA-ANN models (PCA-ANN $^{(1)}$), half of the total data was picked for training (every other seed, 2,612 spectra). One fourth of the total data (1,306 spectra) was kept apart as an early stopping or monitoring set to avoid overtraining of the net, and the remaining data (1,305 spectra) was used for validation. For LWR-PCR models (LWR-PCR $^{(1)}$), the previously spectra used for monitoring and training were joined and used as training set (3918 spectra). The validation set was the same set of spectra used for validation in the PCA-ANN$^{(1)}$ models (1,305 spectra).

The second set of classification models (PCA-ANN$^{(2)}$ and LWR-PCR$^{(2)}$) were trained with data belonging to the first 92 images and validated with the next independent 31 images with fewer RR seeds drawn from a new set of samples not represented in the training set, simulating a real screening situation where the variability from new images and new seeds may not be accounted for in the training set. For PCA-ANN models (PCA-ANN$^{(2)}$), one fourth of the data from the first 92 images was kept apart as a monitoring set (1,305 seeds). Spectra used for all models is shown in TABLE II, after removing outliers.

TABLE II. Spectra (one spectrum per seed) used for the imaging unit models

Discrimination Model	Training Set		Validation set		Monitoring set (PCA-ANN models)	
	Conventional	RR	Conventional	RR	Conventional	RR
PCA-ANN$^{(1)}$	1,283	1,329	633	671	658	648
LWR-PCR $^{(1)}$	1,941	1,977	633	671	----	----
PCA-ANN$^{(2)}$	1,912	2,005	1,544	221	662	643
LWR-PCR$^{(2)}$	2,574	2,648	1,544	221	----	----

(1) Training and validation with seeds from the initial 92 images
(2) Training with seeds from the initial 92 images, validation with seeds from a new set of 31 images, fewer RR seeds from new samples

Single Point Instruments

Instrumentation. Two reflectance instruments were utilized. The Perten DA 7200 (Perten Instruments, Inc., Springfield, IL) is a diode array instrument that covers the wavelength region from 850 to 1,650 nm, taking measurements at 5 nm intervals. A special single seed adapter provided by the company consisting of a concave mirror was used (FIG.3). The instrument was set to take the average of two scans per each seed, taking two blank measurements. The approximate analysis speed was seven seeds per minute; seeds were placed on the concave mirror using tweezers.

The second instrument, called "light tube" in this paper (FIG. 4), was a non-commercial spectrophotometer built by the USDA facility in Manhattan, KS.[10,11] It consisted of a silica tube with 48 miniature tungesten lamps arranged in 6 rows surrounding the tube. A bifurcated fiber optic BIF600-VIS-NIR (Ocean Optics, Dunedin Fla.) with its ends attached to the tube (one in each end) collected the reflected light from the entire seed to be combined in a spectrometer (model NIR256-1.7T1-USB2/3.1/50um SNIR 1074, Control Development, Inc., South Bend, Ind.) that collects from 904 to 1686 nm, at 1nm sampling increments. The blank measurement was taken every 30 minutes as a measurement of the illuminated empty tube. The individual seeds were introduced manually into the tube through a small funnel and a photoelectric switch (D12DAB6FP, Banner Engineering Corp., Minneapolis, MN) located on the top of the tube triggered the spectrometer measurement upon seed detection. Each seed was run through the tube three times and the average from the three scans was taken. This instrument had an approximate analysis speed of fifteen seeds per minute.

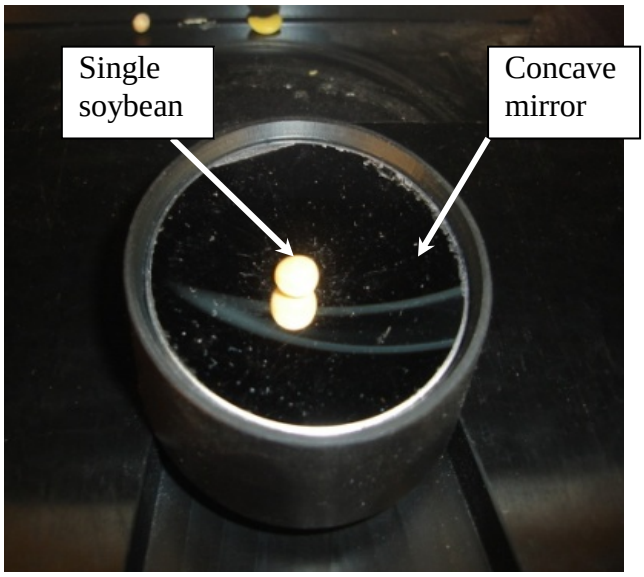


FIG.3. Single seed adapter provided by Perten (Perten Instruments, Inc., Springfield, IL) manufacturers consists of a concave mirror where the seed is placed

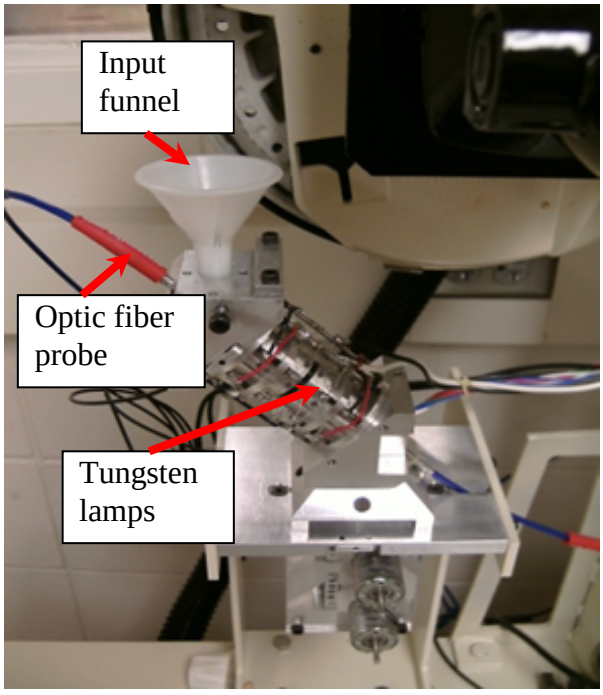


FIG.4. The sampling section of the USDA light tube, with the miniature tungsten lamps surrounding the tube and the funnel where the seeds are introduced.

Samples. The same original samples used in the imaging unit were also used for the single point instrument analysis (total of 227 conventional and 236 RR samples), plus 21 new samples from 2009 crop (6 RR and 16 conventional). Fifteen new seeds were randomly drawn from each sample, 7,275 seeds total. The seeds from each sample were kept in individual bags, not being mixed (keeping the seed identity). The same seeds were scanned in both single point instruments.

Spectra processing. Spectra collected by Perten DA 7200 were reduced to the working range from 955 to 1,645 nm (first and last data point removed) and the final data set had 7,013 spectra. The light tube working range was reduced to wavelengths from 914 nm to 1594 nm. The final data set had 7,274 spectra. Possible outliers were visually detected from the spectra and score plots and removed (only one spectrum for the USDA light tube and around 3% of the Perten data).

Discrimination Models for the Conventional-Single Point Instruments. PCA-ANN$^{(1)}$ models were created including samples in the training set similar to the first models from the imaging unit: ½ of the seeds from all samples were picked as training set, ¼ of the seeds were picked as monitoring set, and ¼ as validation set. All samples were represented in the training set. Perten DA 7200 models were created using 3,504 spectra for training, 1,755 spectra for early monitoring set, and 1,757 for testing. Light tube models had 3,637 train spectra, 1,819 early stop or monitoring set, and 1,819 for validation. LWR-PCR$^{(1)}$ models used the same spectra set as PCA-ANN models for validation, while the monitoring and training set were combined for training.

Fifty eight RR and conventional samples (116 samples total, 1,740 seeds) were kept apart for independent validation of the second set of PCA-ANN$^{(2)}$ and LWR-PCR$^{(2)}$ models. The training sets did not have seeds from the excluded samples. For PCA-ANN$^{(2)}$, again ¼ of the total training spectra were kept apart as an early stop or monitoring test. Table III and table IV show the spectra used for model training and validation, for Perten DA 7200 and the light tube instruments, respectively.

TABLE III. Spectra (one spectrum per seed) used for the Perten DA 7200 models

Discrimination Model	Training Set		Validation set		Monitoring set (PCA-ANN models)	
	Conventional	RR	Conventional	RR	Conventional	RR
PCA-ANN(1)	1,748	1,756	877	880	876	879
LWR-PCR (1)	2,624	2,635	877	880	----	----
PCA-ANN(2)	1,974	1,984	870	869	658	658
LWR-PCR(2)	2,632	2,642	870	869	----	----

(1)Training with seeds from all the samples, validation with seeds from samples represented in the training set

(2)Training with seeds from 369 samples, validation with seeds from new 116 samples

TABLE IV. Spectra (one spectrum per seed) used for the Light Tube models

Discrimination Model	Training Set		Validation set		Monitoring set (PCA-ANN models)	
	Conventional	RR	Conventional	RR	Conventional	RR
PCA-ANN(1)	1,814	1,823	907	911	908	911
LWR-PCR (1)	2,722	2,734	907	911	----	----
PCA-ANN(2)	2,069	2,100	870	855	690	697
LWR-PCR(2)	2,759	2,797	870	855	----	----

(1)Training with seeds from all the samples, validation with seeds from samples represented in the training set

(2)Training with seeds from 369 samples, validation with seeds from new 116 samples

RESULTS AND DISCUSSION

The best classification accuracies achieved are summarized in TABLE V. For LW-PCR, often similar results are achieved with several combinations of neighbors and principal components. The combination giving the highest accuracy requiring the lowest number of PCs is reported.

TABLE V. Summary of best classification results

	Imaging Unit	Light Tube	Perten DA 7200
PCA-ANN (1)			
PCs	10	13	11
Correctly Classified	1,176/1,304 (90.2%)	1,456/1,818 (80.0%)	1,443/1,757 (82.1%)
PCA-ANN (2)			
PCs	10	9	10
Correctly Classified	1,350/1765 (76.5%)	1,327/1,725 (76.8%)	1,386/1,739 (79.7%)
LWR-PCR (1)			
PCs	9	18	18
Neighbors	480	380	440
Correctly Classified	1,160/1,304 (88.9%)	1,713/1,818 (94.2%)	1,625/1,757 (92.5%)
LWR-PCR (2)			
PCs	8	9	8
Neighbors	940	80	540
Correctly Classified	1,370/1765 (74.1%)	1,253/1,725 (72.6%)	1,373/1,739 (78.9%)

(1) Models trained with seeds from all the samples and validation with seeds from samples represented in the training set.
(2) Models trained with 70% of samples and validated with seeds from samples left out. For the imaging unit, models were trained with the first 92 images and validated with the new set of 31 images.

Discrimination by NIRS Imaging

The best PCA-ANN classification accuracies were achieved with nets with 3 hidden neurons and resilient backpropagation training with 20 epochs maximum for the first

models (PCA-ANN(1)) and 15 epochs for the models validated with the new set of 31 images (PCA-ANN(2)). For models validated with seeds from the same images, the classification accuracies were around 90% for both LW-PCR and PCA-ANN algorithms (TABLE V). The number of misclassified seeds was similar for both RR and conventional (TABLE VI). Accuracies dropped below 75% for models created with the spectra from the 92 images and validated with the additional 31 images. Considering that a big part of the conventional seeds belonged to samples represented in the training set, we could conclude that the imaging unit performs the worse when seeds belonging to new images and samples not included in the training set are brought to classification. Most of the misclassifications were conventional seeds classified as RR (TABLE VI).

TABLE VI. Confusion matrix of the classification results from the imaging unit

	PREDICTED							
	PCA-ANN(1)		LWR-PCR(1)		PCA-ANN(2)		LWR-PCR(2)	
ACTUAL	Conv.	RR	Conv.	RR	Conv.	RR	Conv.	RR
Conv.	565	68	578	55	1,134	410	1,159	385
RR	60	611	89	582	5	216	10	211

(1)Training and validation with seeds from the initial 92 images
(2)Training with seeds from the initial 92 images, validation with seeds from a new set of 31 images

The best predictor LW-PCR(2) model required a large number of neighbors (960) which indicates that the imaging unit spectra are noisy. This could be the result from a variety of factors such as non-homogeneous illumination of the sampling surface or low instrument resolution, which add up to the intrinsic diluted signal of imaging technologies when compared to conventional NIRS. The new variability added by new images and seeds from new samples is also reflected in the surface plot of correctly classified seeds for each combination of neighbors and PCs. The surface plot of correctly classified seeds from the LWR-PCR(1) iterations is smoother (FIG.5) compared with the surface plot of LWR-PCR(2) correctly classified seeds (FIG.6).

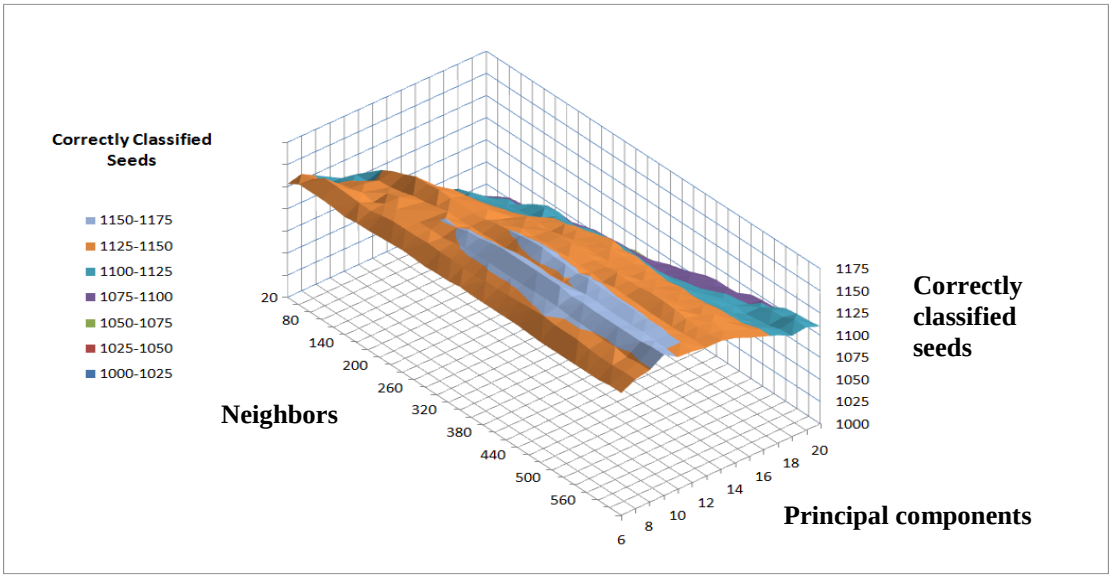


FIG.5. Surface plot of correctly classified seeds from the LW-PCR(1) (validation with seeds from same images and samples of training set) iterations in optimizing the best number of neighbors for the local classification and number of principal components

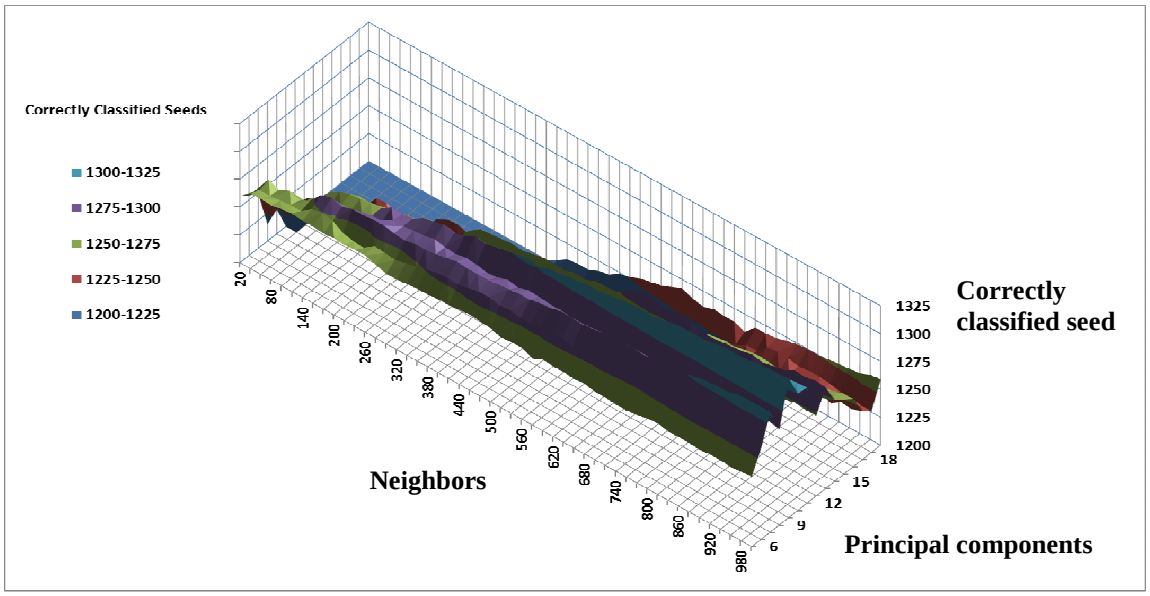


FIG.6. Surface plot of correctly classified seeds from the LW-PCR(2) (validation with seeds from 31 new images) iterations in optimizing the best number of neighbors for the local classification and number of principal components

Discrimination by single point instruments

The best initial PCA-ANN models (PCA-ANN$^{(1)}$), which were validated with seeds from samples already included in the training set, were also obtained with 3 hidden neurons, resilient backpropagation training and around 25 epochs. The best classification accuracies were in the lower 80 percent range for both instruments; those are lower than the ones achieved by PCA-ANN $^{(1)}$ models from the imaging unit (90% correctly classified); this may be due to additional within image variability, which may be modeled and boost the percentage of correctly classified seeds. The best discrimination by LWR-PCR$^{(1)}$ models reached accuracies over 90% for both instruments (TABLE V). The best LWR-PCR models had better accuracies than PCA-ANN; this may not only be due to the intrinsic method approach, but could also indicate the need of better tuning of the neural net parameters. However, the optimization of LW-PCR parameters (neighbors and PCs) is done based on the test set so it increases the risk of overfitting and give optimistic results. Since there are several combinations that give approximately the same accuracies, a second validation test would help determining the best combination among the good ones.

For models validated with seeds from samples kept apart (PCA-ANN$^{(2)}$ and LWR-PCR$^{(2)}$), the accuracies dropped below 80% for both instruments and discrimination methods. The accuracy decrease specially impacted the amount of RR misclassified in LWR-PCR models (TABLE VII and TABLE VIII), which resembled the results from PCA-ANN$^{(2)}$ for Perten DA 7200. LWR-PCR$^{(2)}$ model with light tube data lead to an amount of misclassified RR seeds three times higher than the amount of misclassified conventional seeds (TABLE VIII). Perten DA 7200 LWR-PCR$^{(2)}$ models showed slightly better classification accuracies compared to the light tube. Light tube spectra are visibly affected by seed size and shape (FIG. 7). This characteristic may help boosting the LWR-PCR model accuracy when seeds belong to samples represented in the training set, as additional information regarding seed physical traits is already represented in the training set. But on the opposite, if the seeds do not belong to any pre-existing sample in the training set, the resemblance to other spectra is low. This is reflected in TABLE V, where

LWR-PCR$^{(1)}$ for the light tube achieves accuracies of 94%, but for LWR-PCR$^{(2)}$ the optimum number of neighbors is much smaller than Perten 7200 and the imaging unit, with classification accuracies significantly lower than Perten DA 7200. The difference among instruments was smaller for PCA-ANN $^{(2)}$ models.

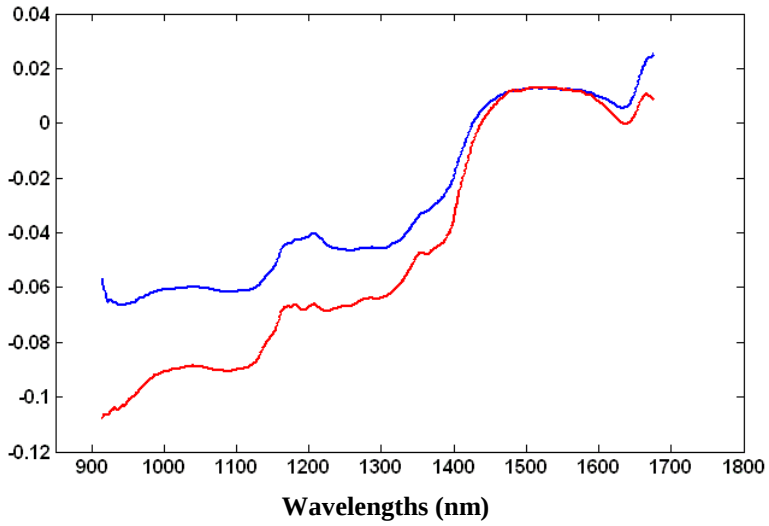


FIG.7. Light tube spectra of two soybean seeds from two different samples. The difference among them is not only visually detected on baseline differences but on specific peaks.

TABLE VII. Confusion matrix of the classification results from the Perten DA 7200

	PREDICTED							
	PCA-ANN$^{(1)}$		LWR-PCR$^{(1)}$		PCA-ANN$^{(2)}$		LWR-PCR$^{(2)}$	
ACTUAL	Conv.	RR	Conv.	RR	Conv.	RR	Conv.	RR
Conv.	781	96	825	52	754	116	745	125
RR	218	662	80	800	237	632	241	628

(1) Validation with seeds from samples included in the training set
(2) Validation with seeds from 116 independent samples

TABLE VIII. Confusion matrix of the classification results from the Light Tube

	PREDICTED							
	PCA-ANN(1)		LWR-PCR (1)		PCA-ANN(2)		LWR-PCR(2)	
ACTUAL	Conv.	RR	Conv.	RR	Conv.	RR	Conv.	RR
Conv.	751	156	874	34	768	102	794	76
RR	206	705	71	839	296	559	396	459

(1) Validation with seeds from samples included in the training set
(2) Validation with seeds from 116 independent samples

CONCLUSIONS

In this study we demonstrated that discrimination at a single seed level of conventional and RR soybean is possible using Near Infrared reflectance technologies and non-linear models such as PCA-ANN and LW-PCR. Similar accuracies to the ones achieved by transmittance of bulk samples[7] were attained whenever the seeds to be classified are represented in the training set. PCA-ANN(1) models showed lower accuracies in single-point instruments, probably indicating that either there may be variability within images from the imaging unit which has been accounted during the training and favored the classification of seeds included in the same images of training seeds, or that the two single point instrument fell in a local minimum during the training. The classification accuracies among single-point technologies were not very distant from one to another, although the light tube combined with LWR-PCR models showed the most promising performance because the instrument is more susceptible to seed physical characteristics. That helped classifying seeds with good accuracies whenever the sample is represented in the training set. On the opposite, this characteristic negatively impacts the accuracies if an unknown seed belongs to a sample not represented in the training set. The imaging

unit performed slightly worse. The accuracy of models validating with seeds represented in the training sets from coming from new images has not been evaluated, but the new set of 31 images which had more than half of seeds from samples already represented in the training set lead to accuracies close to the ones from single point instruments, validated with seeds entirely coming from new samples. From the single-point instruments, a highest proportion of misclassified seeds were RR classified as conventional. For the imaging unit, it was the opposite. This last would be preferred, since highest sensitivity to RR could serve as a first screening precaution, and any sample not passing the screening could be later reanalyzed by another official method.

Although the attained accuracies (higher 80 - lower 90 percent range) are not high enough to allow using these technologies as a solo discrimination tool under most of the restrictive legislative thresholds, NIRS could be used as a screening method for farmers and elevators because it is low cost and does not require special sample preparation. Technologies such as the USDA light tube allow fast scanning of whole seed batches without the need of subsampling, thus reducing this error associated with current official GMO detection methods. Future application of the method could focus on comparing the NIRS discrimination ability of entire seed batches versus the subsampling error using traditional methods. Any correlation could help creating strategies for RR detection, either using NIRS alone or combining NIRS with an official method, improving the overall detection/discrimination process. Locally weighted algorithms are the best overall choice according to current results. The method requires fewer parameters to be optimized and may lead to better accuracies. However, a careful process for optimizing the number of neighbors and PCs used in the local models should be carried out. Partial least squares (PLS) regression could be tried to obtain more parsimonious local models.[12]

It has been shown that although the current discrimination models included a wide range of samples (harvesting years, overall composition) in the training set, those are still not enough to attain good accuracies when new seeds from new samples must be discriminated. Having over 150 samples of each class represented in the training set was not enough to classify seeds from new samples with good accuracies.

REFERENCES

1. S. Konduru, J. Kruse, and N. Kalaitzandonakes, “The Global Economic Impacts of Roundup Ready Soybeans”, in *Plant Genetic/ Genomics, Genetics and Genomics of Soybeans*, G. Staycey, Ed. (Springer, New York, NY), vol.2, part IV, p.375.
2. H. M., Hanna, G. R., Quick, and D. H. Jarboe, Iowa State University Extension publications. Retreived March 2006 from: www.extension.iastate.edu/Pages/ grain/publications/grprod/02icmm.pdf
3. Friends of the earth, briefing resource (2006). Retrieved January 2011 from: http://www.foe.co.uk/resource/briefings/gm_animal_feeds.pdf
4. J. Miljuš-Djukić,B. Banović, Z. Jovanović, D. Majić, M. Milisavljević, J. Samardžić, and G. Timotijević, Romanian Biological Lett**. 15(1)**, 102 (2010).
5. A. Rizzi, C. Sorlini, and D. Dalfonchio, AgBiotech. Net. **6**, 1N (2004).
6. P. Hübner, H.-U. Waiblinger, K. Pietsch, and P. Brodmann, J. AOAC Int. **84 (6)**, 1855 (2001).
7. S. A. Roussel, C. L. Hardy, C. R. Hurburgh Jr., and G. R. Rippke, Appl. Spectrosc. **55(10)**, 1425 (2001)
8. P. Geladi, D. MacDougall, and H. Martens, Appl. Spectrosc. **39**, 491 (1985).
9. A. A. Gowen, C. P. O’Donell, J. M. Frias, G. Downey, First workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. WHISPERS '09,1 (2009).
10. P. R. Amstrong, Appl. Eng. Agric. **22**, 767 (2006).
11. J. G. Tallada, N. Palacios-Rojas, and P. R. Armstrong, J. Cereal Sci. **50**, 381 (2009).
12. V. Centner and D. L. Massart, Anal. Chem. **70**, 4206 (1998).

CHAPTER 3. DISCRIMINATION OF CONVENTIONAL AND ROUNDUP READY SOYBEAN SEEDS. TRANSMITTANCE VS REFLECTANCE MEASUREMENTS AND MOISTURE EFFECT

A paper to be submitted to the Journal of Near Infrared Spectroscopy

Lidia Esteve Agelet, Glen R. Rippke, Charles R. Hurburgh*

Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011

* To whom correspondence should be addressed at 1547 Food Science Building, Iowa State University, Ames, IA 50011-1060. Phone: 515-294-8629, Fax: 515-294-6383, Email: tatry@iastate.edu

ABSTRACT

Roundup Ready® soybeans which have been genetically modified to be resistant to Roundup® herbicide, is one of the first genetically modified crops recognized safe and commercialized. However, most of the current worldwide regulations for importing and exporting food demand the control, identification, and proper labeling of all genetically modified agriculture products. Previous studies showed that Near Infrared Spectroscopy (NIRS) could distinguish among Roundup Ready® and conventional soybeans at bulk and single seed sample level (classification accuracies between 80 and 94%). In this paper we focus on single seed discrimination of fewer conventional varieties (five) with their respective Roundup Ready® version. Analysis carried out in Fourier Transform transmittance mode and whole-surface reflectance in single seed levels lead to better discriminations from reflectance mode with either using Least Squares Support Vector Machines (accuracy of 82%) or Locally Weighted Principal Component Regression

(accuracy above 90%). Varieties with higher misclassification rates had the lowest bulk moistures, but when validating the models, seeds at higher moistures all were classified as Roundup Ready®. That indicates a possible interaction between absorption of carbohydrate bonds relevant in the discrimination models and moisture. The excellent discrimination accuracies within varieties (above 95% for most varieties) showed once more how the Roundup Ready® gene generate changes in the seeds which make them different from the conventional and are easily measurable by NIRS. This could help breeders to obtain lots with higher purity of either conventional or Roundup Ready® soybean seeds.

KEYWORDS: Roundup Ready®, soybeans, transmittance, moisture, discrimination

INTRODUCTION

Genetically modified (GM) organisms have been manipulated to avoid diseases, to increase resistance to herbicides, and to increase their nutritional value. Their worldwide acceptance has been accompanied by controversy regarding their safety for humans (with introduction of new allergens and resistance to antibiotics) and environment (biodiversity issues).[1,2] For this reason, most countries set regulations for identification, quantification, and appropriate labeling of products containing genetically modified organisms (GM). Roundup® is a popular glyphosate-based herbicide produced by Montsanto Company; Roundup® kills a broad variety of weeds on contact. Roundup® application in fields used to be only adequate to crops at certain development stages and direct application had to be avoided.[3] The development of crops with resistance to the herbicide reduced those hassles and restrictions. The patent and marketing of Roundup® resistant crops, licensed with the name of Roundup Ready®, was done by Montsanto. By genetic recombinant DNA technology, genetic material from the bacteria *Agrobacterium Tumefacien* was introduced to the crop genoma, conferring the crop a high tolerance to the herbicide, leading to less restrictive use of Roundup®, lower production costs, and even higher crop yields.[4]

Soybeans (*Glycine max L.)* were the first Roundup Ready® (RR) crop to be introduced in the markets in 1996. They rapidly displaced conventional soybeans for the previously mentioned advantages on crop management and yields, being RR currently more than half of the soybean fieldcrops around the world.[5] RR soybeans are widely accepted in the global markets; they are one of the two currently accepted GM varieties of soybeans in Europe, which has the most restrictive laws regarding GM importation. But despite of their acceptance, they must be labeled as a GM crop, even if they are present as adventitious contamination in conventional batches whenever their percentage exceeds pre-established thresholds, determined in function of several parameters such as company and consumer requests or political aspects.[6] Current thresholds of adventitious GM contamination in conventional soybeans for feeding range from 0.9% (i.e Europe) to 5% (i.e. Japan or Taiwan). In the case of Europe, the tolerance limit applies to contamination of recognized GM varieties; otherwise, the threshold is reduced to 0.5 % whenever the varieties are proved safe.

Determination of GM traces in big batches is challenging. On one hand, current determination methods are time consuming and complex, not suitable for on-site measurements (laboratory-based methods). Analyses are divided as Protein-based methods and DNA-based methods. Protein-based methods such as Enzyme-Linked Immuno Sorbent Assay (ELISA) work with specific antibodies, which require previous knowledge of the GM to be analyzed, but are fastest, cheapest, and simpler than DNA-based methods.[6] DNA-based methods, such as Polymerase Chain Reaction (PCR), are more sensitive: the lowest limit of detection of GM DNA material is around 0.1%.[7] On the other hand, these methods are destructive, meaning that even in the case they could perform faster and be applied on-line, only a small portion of the sample could be analyzed. This leads to the added problem of taking representative samples from big batches, proved to be in function of the type of grain and the threshold to be analyzed.[8]

The first attempt to use Near Infrared (NIR) to discriminate RR and conventional soybeans was done by Roussel et al.[9] In their study, NIR transmittance was used to scan over 4,000 bulk soybean samples from each class. Non-linear classification methods such as locally weighted principal component regression (LW-PCR) and artificial neural

network (ANN) were required to achieve classification accuracies of 93 and 88% respectively, using independent validation sets. A more recent study[10] was carried out to study the feasibility of NIR reflectance to discriminate among RR and conventional single seeds, and thus be used as method to determine seed lots impurity. Three reflectance NIR technologies were used: a chemical imaging unit, a single point instrument with a single seed adapter, and the USDA light tube.[11,12] Over 5000 seeds belonging to around 240 samples (15 seeds per sample) of each class (RR and conventional) were scanned. LW-PCR accuracies were best for the light tube (94%) when seeds for validation were picked from samples represented in the training set. ANN results were similar for all technologies, with accuracies in the lower 80% range. Those results were very similar to those of Roussel et al.[9] using independent validation sets. When seeds were from samples not represented in the training set, regarding that models were developed with over 150 samples from each class, classification accuracies dropped in all the instruments. The best accuracy was achieved by the traditional single point instrument (79%).

In this paper we use a Fourier-Transform Near infrared Transmittance (FT-NIR) and the USDA light tube instruments.[11,12] Transmittance measurements have higher throughputs and may be more accurate when analyzing heterogeneous samples, as the irradiated light goes through the entire sample. The USDA light tube, on the other hand, takes reflectance spectra from whole seeds. Both measurement modes are compared. A smaller set of 10 samples, 150 seeds picked per sample, were tested. Five conventional varieties and their five corresponding varieties with the RR gene were utilized. The objectives of the study were (1) to compare discrimination accuracies between the whole seed reflectance mode by the light tube and FT transmittance measurements in discrimination of RR soybeans and conventional, (2) to compare the performance of two different classification algorithms: Least Squares Support Vector Machines (LS-SVM) and Locally Weighted Principal Component Regression (LW-PCR) (this last proven to outperform ANN in the previous research), (3) compare accuracies with the previous study with more samples and fewer seeds per sample, (4) to determine and compare the discrimination accuracies within a single variety (conventional and RR modified) and

when more than one variety is involved, and (5) to analyze the impact of seed moisture on discrimination accuracy, using higher moisture seeds for validation of the obtained models. Moisture was suspected to have influence in discriminating bulk samples by transmittance. [9] Since the application of the method in elevators or commodities suppose variable seed moistures, any effect that moisture may induce to the discriminative ability of the models should be acknowledged.

MATERIAL AND METHODS

Samples

Five conventional public soybean varieties from 2007 crops (labeled as M97-302, M97-303, M97-304, M97-305, and M97-306 varieties) and the same respective varieties with the Roundup Ready® (RR) gene transferred were used in this study. This made a set of 10 samples. A hundred and fifty seeds from each sample were picked and scanned by the two instruments consecutively (1,500 scanned seeds total). The initial average moisture of the bulk seeds was measured by scanning them with a Infratec 1221 transmittance instrument (Foss North America, Eden Prairie, MN, USA) using a cuvette and the Iowa State moisture calibration. Sample composition predicted with NIRS Iowa State calibrations is shown in table 1. The standard error of prediction (SEP) of moisture calibration was 0.37%, SEP=0.52% for protein, SEP=0.37% for oil, and SEP=0.08% for fiber. No appreciable physical differences were detected among the conventional and RR samples within a variety. Additional sets of 150 seeds more from each sample were picked and sealed in individual small plastic bags with a wet paper towel on the top. The bags were kept in the fridge for around 3 weeks or until their average moisture was over 13%. The moisture was monitored and predicted with the Infratec 1221 instrument. During that period of time, the paper towels were replaced when dry and seeds were shaken to allow better distribution of the moisture within samples. After scanning each seed in the two instruments consecutively, the approximate moisture of each seeds was

estimated by weight difference oven drying for 3 hours at 130 C (AOCS, Ac 2-41 method).

Table 1. Composition of the ten samples used in the study, predicted with NIR transmittance with Iowa State calibrations

Sample	Initial Moisture (%)	Protein (%)*	Oil (%)*	Fiber (%)*
M97-302 RR	8.8	34.9	16.9	5.0
M97-302	8.6	36.1	18.1	4.8
M97-303 RR	8.4	36.0	18.8	4.7
M97-303	8.4	36.4	17.4	4.8
M97-304 RR	8.2	37.9	17.0	4.7
M97-304	8.3	36.2	18.0	4.8
M97-305 RR	9.3	38.0	17.3	4.6
M97-305	8.9	36.3	17.9	4.8
M97-306 RR	9.5	36.2	18.3	4.7
M97-306	8.9	34.6	18.0	4.7

* 13% moisture weight basis

Instrumentation

In this study, two spectrometers were used. Buchi NIRFlex N-500 (Buchi Corporation, New Castle, DE) is a Near Infrared Fourier Transform (FT-NIR) spectrometer which has the capability of working in both reflectance and transmittance mode with the use of appropriate modules. For this research, the NIRFlex solids transmission module with the 10-well sample cell (Figure 1) was used to analyze the individual seeds. The instrument covers the spectral range from 11,520 to 6,000 cm-1 (868.1 – 1,666.7 nm), with 4 cm-1

sampling increment (1,381 data points) at full resolution of 8 cm-1. The second spectrometer (Figure 2) was designed and built by the United States Department of Agriculture (USDA) in the Manhattan (KS) facility and differs from conventional single point spectrometers in the fact that takes spectra from all the seed surface. The light tube instrument sample cell is a silica tube where a single seed passes through being illuminated by 48 miniature tungsten lamps located surrounding the tube. A photoelectric switch D12DAB6FP (Banner Engineering Corp., Minneapolis, MN) detects when the seed is manually dropped to the tube from a small funnel located on the top of the tube and the instruments starts the reflectance spectra measurement from the whole seed through Y-shaped bifurcated fiber optic BIF600-VIS-NIR (Ocean Optics, Dunedin Fla.) with two of the ends attached one in each end of the tube. The third end of the fiber optic is connected to the CDI spectrometer model NIR256-1.7T1-USB2/3.1/50um SNIR 1074 (Control Development, Inc., South Bend, Ind.) with 1 nm sampling increments from 904 to 1686 nm. More information related to the instrument can be found in the literature. [11,12] The blank measurement, taken as a measurement of the empty illuminated tube with the funnel, was taken every 20 minutes. Each seed spectrum was the average of three spectra.

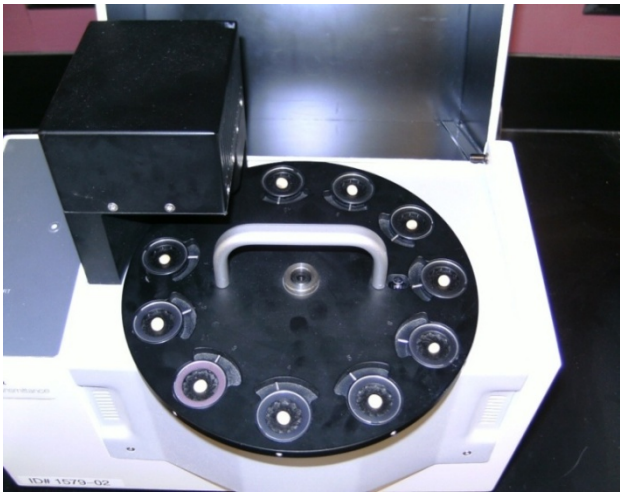


Figure 1. FT-NIR Buchi NIRFlex N-500 working in transmittance mode with the 10-well sample cell

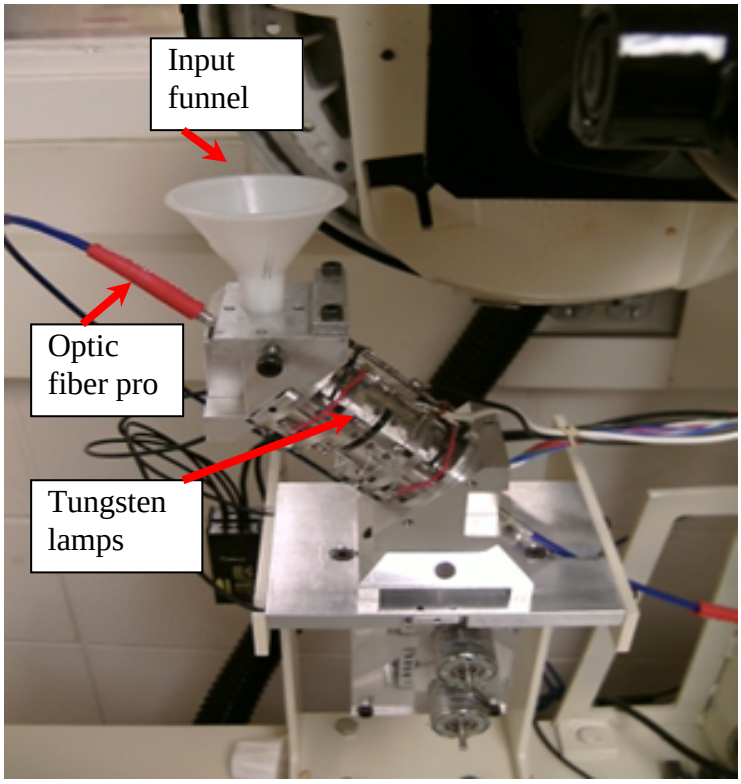


Figure 2. Image of the tube surrounded by the miniature tungsten lamps with the funnel, where seeds are introduced, on the top. The two extremes of the Y-shaped optic fiber are connected on the top and on the bottom of the tube.

Data Management and Discrimination Models

Spectral data was used raw (as apparent absorbance measurement), as previous tests indicated that spectral pretreatments such as standard normal variate (SNV) or Savitzky-Golay derivatives decrease classification accuracies, probably because they remove relevant information. For both instruments, wavelengths from the two extremes of the spectra were removed. For the FT-NIR instrument, 10 data points were removed from each extreme. The working wavelength range was from 6,040 cm-1 to 11,480 cm-1 (871 – 1655 nm). The working spectra from the light tube covered the range from 953 nm to

1636 nm (50 data points removed from each extreme). Data was imported to The Unscrambler v.9.8 (Camo AS, Trondheim, Norway). Detection and outlier removal tasks were carried out both visually and with principal component analysis (PCA) within varieties. Samples showing either extreme scores in the first 14 principal components (PCs) or having high leverage vs residual values when compared with the rest were flagged as possible outliers. Discrimination models were developed in Matlab 7.10.0 (2010a) (Mathworks, Natick, MA). Two classification algorithms were tested: least squares support vector machines (LS-SVM) and locally weighted principal component regression (LW-PCR).

LS-SVM models were developed using LS_SVMlab v.1.5 toolbox functions.[13] We used the Guassian Radial Basis Function (RBF) as the kernel function for high-dimensional mapping (non-linear discrimination). Both kernel width (σ) and the model regularization parameter (γ) were simultaneously tuned using grid search based on 15-block cross-validation for the general model, and 10 block cross-validation for within-variety models. The inputs were the PCA scores after autoscaling the spectra, testing from 9 to 15 principal components. The number of optimum PCs were selected according to the best classification results in the external validation set. PLS_toolbox v.3.5.4 (Eigenvector Research Inc., Wenatchee, WA) functions were used for data processing and PCA, while the original algorithm for LWR-PCR was retrieved from the PLS toolbox v.2.1.1 (Eigenvector Research Inc., Wenatchee, WA). In that algorithm, the new sample is assigned to a class (either 0 or 1) according to a local model built with its neighbors. The neighbors are samples which have resemblance or closeness to the new sample spectra, measured by Mahalanobis distance in the first principal component. The relevance of each neighbor in the PCR classification of the new sample is defined by a cubic weight function (Equation 1) which is function of d, the Mahalanobis distance. Both numbers of neighbors and PCs for the PCR discrimination were optimized by iteration, considering the best combination the one that lead to a small number of misclassifications when predicting the validation set.

Equation 1. $W_S = \left(1 - d^3\right)^3$

Experimental Design

General classification models which included all dry seeds (table 1) from the 5 varieties (both RR and conventional) were created with data from both instruments (FT-NIR transmittance and light tube reflectance) and both discrimination models (LS-SVM and LWR-PCR). Two thirds of the data was used for training (approximately 1000 spectra), and one third was kept for validation (500 spectra). Five other additional classification models were created to discriminate within single varieties (conventional versus RR). In each single-variety model, also two thirds of the data was used for training (200 spectra) and one third for validation (100 spectra). The experimental design with the final number of samples is summarized in table 2. Outliers were only evident from the FT-NIR data; no clear outliers were spotted in the light tube instrument.

Table 2. Used spectra in both discrimination algorithms

		Varieties	Training Spectra		Validation Spectra	
			Conv.	**RR**	**Conv.**	**RR**
Overall Classification	**Buchi FT-NIR**	5	485	454	250	250
	Light tube reflectance	5	500	500	250	250
Within-Variety Classification	**Buchi FT-NIR**	1	(1) 99	(1) 98	(1) 48	(1) 50
			(2) 100	(2) 97	(2) 48	(2) 49
			(3) 93	(3) 98	(3) 50	(3) 49
			(4) 98	(4) 97	(4) 50	(4) 50
	Light tube reflectance	1	100	100	50	50

(1) M97-302, (2) M93-303, (3) M97-304, (4) M97-305, (5) M97-306

Moisture Effect

Both LS-SVM and LW-PCR general models were validated with high moisture seeds. The first validation was carried out using seeds from a single variety (M97-304, later shown to be variety with higher misclassifications). The tested seeds belonged to different moisture ranges, in order to determine if there was a moisture threshold in which classifications changed. A second validation set included seeds from all the varieties, covering a wider high moisture range (13.5 – 17%) (Table 4). The amount of seeds from each test set was limited by the moisture achieved by each individual sample brought to moisture increase. The final moisture was not homogeneous, and differed between and within samples.

Table 3. Composition of the validation set including seeds from the variety M97-304

Moisture (%)	Conv.	RR	Total
8.5 - 10 %	3	4	**7**
10 – 13.5 %	16	12	**28**
13.6 – 15 %	18	26	**44**

Table 4. Composition of the validation set including all varieties with high moisture range (13.5 – 17%)

	Varieties					
	M97-302	**M97-303**	**M97-304**	**M97-305**	**M97-306**	**Total**
Conv.	20	20	20	20	20	**100**
RR	17	20	20	20	10	**87**

RESULTS AND DISCUSSION

Overall Classification with Dry Samples

FT-NIR Instrument. The best classification accuracy applying LS-SVM (392 correctly classified spectra over 500 test spectra, 78%) was achieved using 13 PCs. According to the number of spectra that had to be deleted, the instrument shows problems related with sample positioning which may be added to the variability caused by differences in seed thickness and light pathlenght. This sensitivity negatively influenced the spectra and thus the discrimination model. Figure 3 shows the number of misclassified seeds of each variety from the validation set. The RR samples of varieties M97-303 and M97-304 had the highest misclassifications, while variety M97-305 had the lowest. Both M97-303RR and M97-304RR had the lowest bulk moisture predictions of the entire set (table 1). M97-305RR and M97-306RR are the two samples with highest bulk moisture, and got the lowest misclassification results from the RR samples.

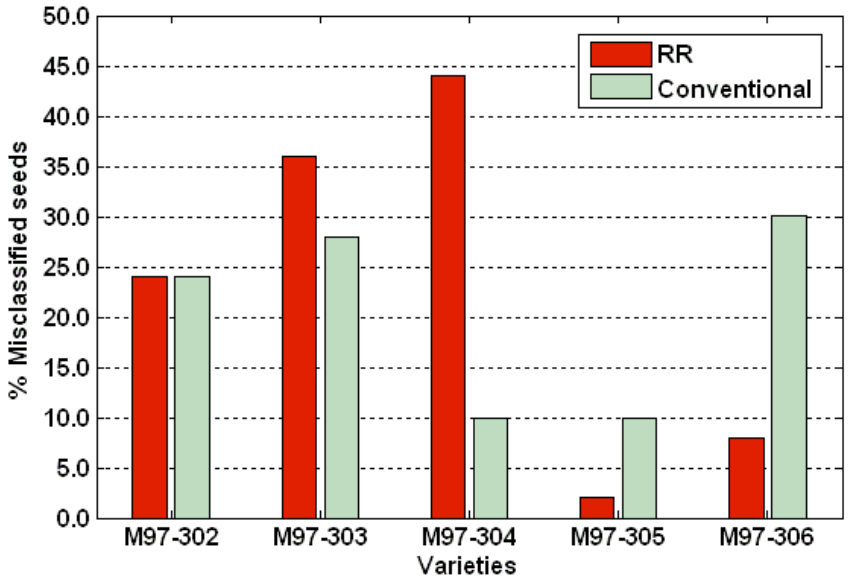


Figure 3. FT-NIR transmittance Buchi instrument misclassifications per variety from the validation of the LS-SVM model

Higher discrimination accuracies were achieved with the use of LW-PCR. The correctly classified rate was 88% (443 correctly classified over 500 test spectra) with a number of neighbor samples of 380 and 11 PCs used in local models. This result is similar to the achieved by the reflectance imaging unit from an earlier study.[10] Noise originated by sample positioning overcome the advantage of high throughput of FT-NIR transmittance, leading to results close to what a low resolution reflectance imaging unit gave. Misclassifications from the validation set are shown in figure 4. The 57 misclassified seeds followed a different pattern than the ones misclassified by LS-SVM. Some samples (e.g. M97-304RR) still were problematic to be discriminated from conventional seeds, most of the RR tested samples achieved better discrimination accuracies compared with the LS-SVM model. Compared with LS-SVM misclassifications, moisture did not seem to be correlated with LW-PCR misclassifications.

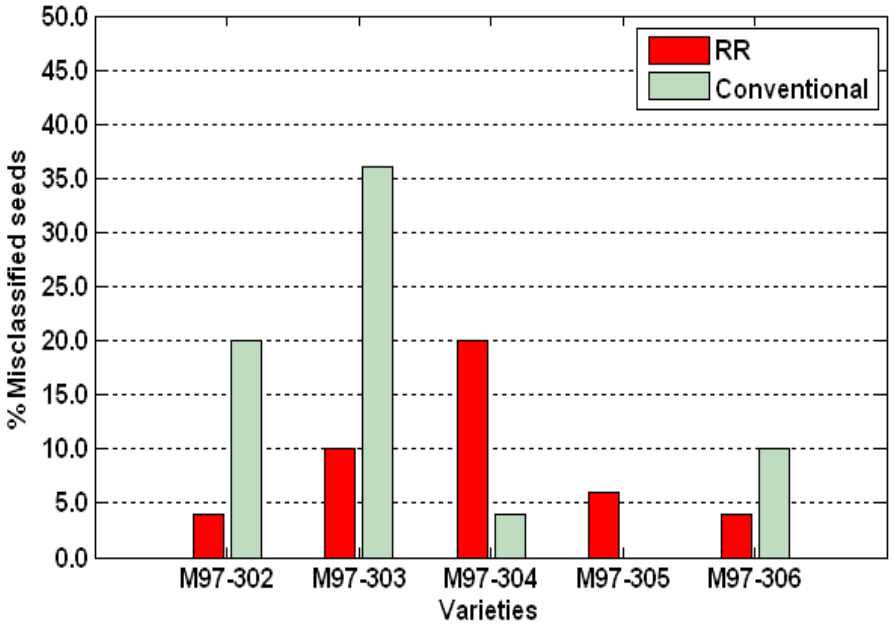


Figure 4. FT-NIR transmittance instrument misclassifications per variety, using the optimum LW-PCR overall classification model

Light Tube Instrument. The instrument outperformed the classification accuracies obtained by the FT-NIR instrument. LS-SVM validation achieved 411 spectra correctly classified over 500 (82%), using 14 PCs. This result is similar to the previous study[10] with PCA combined with artificial neural networks (ANN) models were developed using a larger number of samples represented by a smaller number of seeds. We could conclude that algorithms such as LS-SVM and ANN may not lead to better accuracies when they are trained with seeds from fewer samples, and being each sample represented by a high number of seeds.

Figure 5 shows the bar plot of the validation set misclassifications from each variety. The highest misclassification rates were from M97-304RR and M97-305 conventional. However, the third sample with highest misclassifications was M97-303RR. M97-305RR and M97-306RR had the lowest misclassification rates, and the highest bulk moistures (Table 1). This agrees with previous suspicions[9] of moisture having influence on the classification. RR varieties with higher bulk moisture seem to have better classification rates in LS-SVM models, and the opposite.

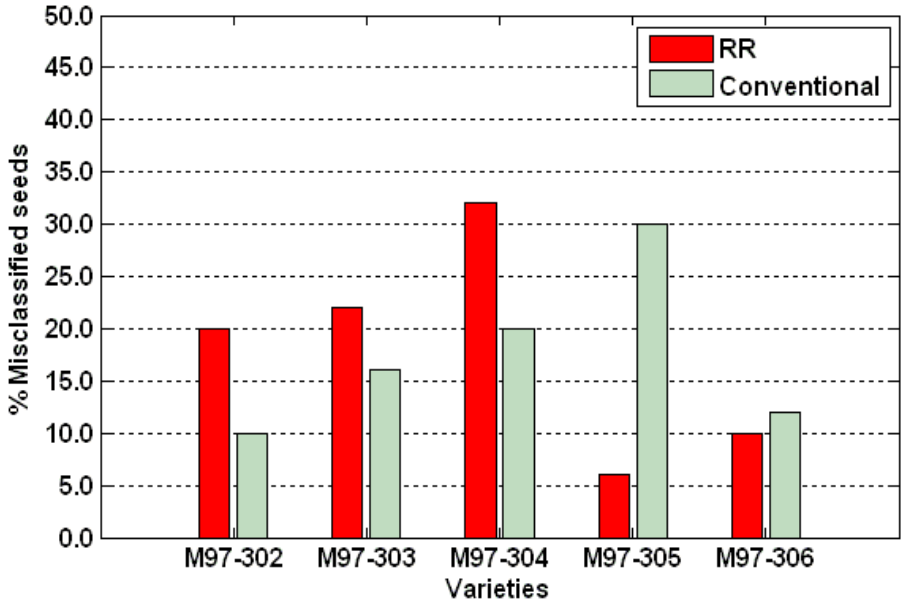


Figure 5. USDA reflectance light tube instrument misclassifications per variety from the validation of the LS-SVM model

With LWR-PCR, the validation set could be discriminated at 99% (494/500) using a smaller number of neighbors (240) and 14 PCs. Several combinations lead to similar accuracies. Higher accuracies can be achieved for local models with 190 – 300 neighbors and 18 PCs (499 correctly classified spectra over 500, 99.8%), but there is a higher risk of overfitting. Those accuracies exceed the ones also reported for LWR-PCR in a previous study using the light tube instrument (94%).[10] As opposted to LS-SVM and ANN, LW-PCR models may benefit of having more seeds representing each sample in the training set. However, it should be taken in account that LS-SVM regularization parameter and kernel width have been optimized by cross-validation (the optimum number of PCs was optimized from the validation set), while optimal number of neighbors and PCs for local models in LW-PCR was determined from the validation set. This can result in slightly optimistic results for those models.

Within Variety Classification

Within variety classifications accuracies with LS-SVM were overall much better than the general model with all the varieties (table 5), except of those which already had higher misclassification rates in the previous validation (e.g. M97-305). As seen in table 1, some of the varieties have slight differences in bulk protein or oil content between RR and conventional, but it cannot be considered that the discrimination among them is due to differences in concentrations from any of those compounds.

Table 5. Validation results of within variety LS-SVM models

Instrument	Variety	Pcs	Misclassified		Total Accuracy (%)
			Conv.	RR	
Buchi FT-NIR Transmittance	**M97-302**	8	1	2	96.8
	M97-303	12	7	17	75.3
	M97-304	15	9	9	81.8
	M97-305	12	4	1	95.0
	M97-306	13	3	4	92.7
USDA Light Tube Reflectance	**M97-302**	13	8	1	91.0
	M97-303	13	2	7	91.0
	M97-304	10	14	13	73.0
	M97-305	13	3	1	96.0
	M97-306	10	4	5	91.0

LW-PCR models within each variety were again better than LS-SVM as shown in table 6. Most of the varieties can be discriminated at 100% of accuracy with the USDA light tube and over 90% with the FT-NIR Buchi. The average of neighbor points needed for local models range from 25 to 55. Close attention when using such a small number of neighbors should be paid; especially if the amount of required PCs is relatively large. This could be the case of local models from varieties such as M97-302 and M97-304 from the light tube data. For most of the varieties, using a lower number of PCs (5,6) and around 50 neighbors still lead to accuracies over 90%. However, good accuracies can be achieved using other combinations of neighbors and PCs, which could be tested with a second validation set.

Table 6. Validation results of within variety LW-PCR models

Instrument	Variety	Pcs	Neighbors	Misclassified		Total Accuracy (%)
				Conv.	RR	
Buchi FT-NIR Transmittance	**M97-302**	9	55	0	0	100.0
	M97-303	13	165	0	6	93.8
	M97-304	10	105	0	1	99.0
	M97-305	5	15-80	0	1	99.0
	M97-306	5	35-80	0	0	100.0
USDA Light Tube Reflectance	**M97-302**	8	25	0	0	100.0
	M97-303	9	55	0	0	100.0
	M97-304	13	55	1	0	99.0
	M97-305	8	35	0	0	100.0
	M97-306	8	45	0	0	100.0

Discrimination of High Moisture Seeds

The validation of the general model carried out by seeds from the M97-304 variety at different high moisture ranges lead to a completely opposite misclassification pattern for this variety when compared with the validation results done with dry seed spectra. Most of the RR seeds were correctly classified and most of conventional seeds were misclassified as RR. The results were the same for both LS-SVM and LW-PCR algorithms. The misclassifications were similar among moisture rates, so a slight increment of moisture leaded to the same results as higher moisture increments.

Table 7. Validation of LS-SVM models from both instruments with seeds from variety M97-304. Misclassified seeds over total per moisture range.

	Transmittance NIR-FT		Light Tube Reflectance	
Moisture (%)	**Conv.**	**RR**	**Conv.**	**RR**
8.5 - 10 %	3/3	1/4	2/3	0/4
10 – 13.5 %	14/16	1/12	15/16	2/12
13.6 – 15 %	17/18	2/26	15/18	0/26

Table 8. Validation of LW-PCR models from both instruments with seeds from variety M97-304 RR. Misclassified seeds over total per moisture range.

	Transmittance NIR-FT		Light Tube Reflectance	
Moisture (%)	**Conv.**	**RR**	**Conv.**	**RR**
8.5 - 10 %	3/3	0/4	1/3	0/4
10 – 13.5 %	15/16	0/12	11/16	0/12
13.6 – 15 %	17/18	2/26	15/18	0/26

Similar results were achieved when validating the LS-SVM general models with high moisture seeds from all the varieties (Figure 6). For models from both instruments, most of the conventional seeds were classified as RR. RR samples showed lower misclassification rates, although samples with higher misclassifications were not the same for both instruments. These results show that the hyperboundary created with LS-SVM algorithm to discriminate conventional and RR seeds was relying on the information located in the water absorption region, which is located also in the carbohydrate absorption. The validation of LW-PCR models with high moisture seeds lead to different results for both instruments. While the LW-PCR predictions for FT-NIR also had all the RR seeds correctly classified and all conventional misclassified as conventional as what happened with LS-SVM model validation, the predictions from the light tube were random. High misclassification rates were observed for both RR and

conventional samples (Figure 7). This could indicate that the combination of neighbors and PCs in the last model may not be the optimal but could be overfitted to the first validation set, focusing on modeling intrinsic set features instead of information which allows generalization of the discrimination ability.

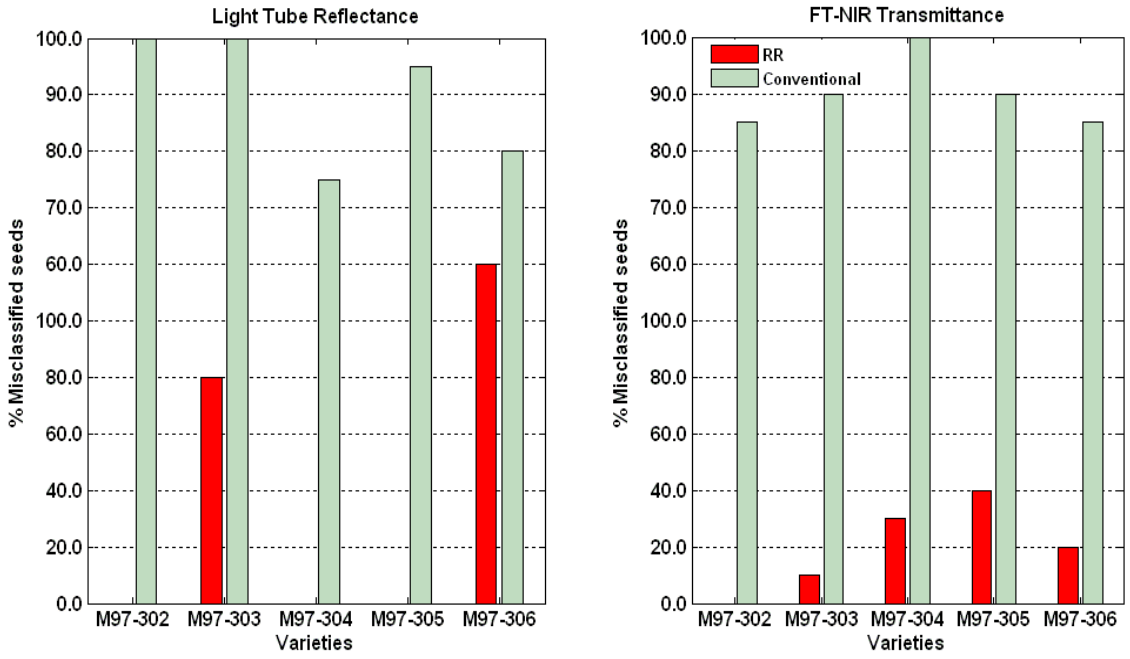


Figure 6. Misclassified seeds from the validation of the initial LS-SVM models (Light tube model to the left and FT-NIR model to the right) with high moisture seeds (13.5-17%) from all the varieties.

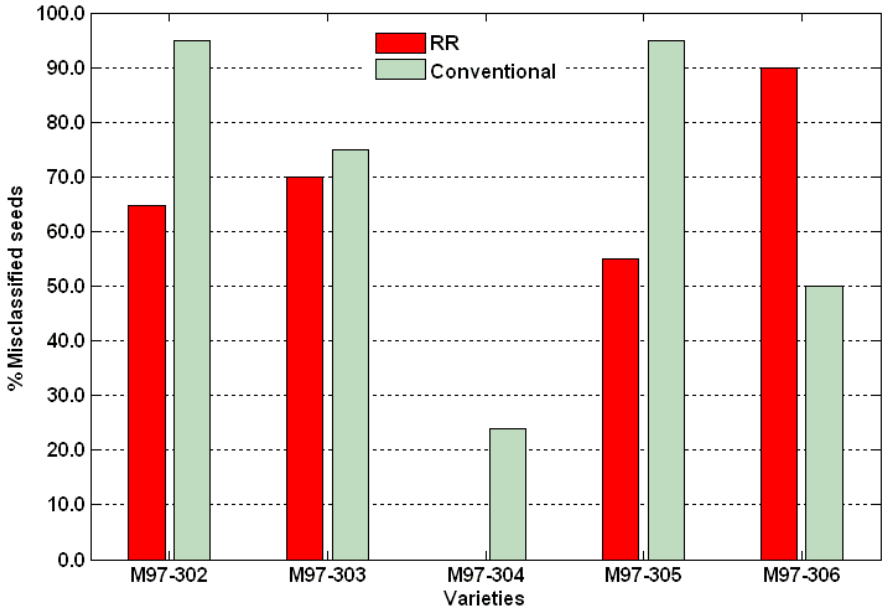


Figure 7. Validation of the Light Tube LW-PCR model with high moisture seeds (13.5-17%) from all the varieties

The initial guess from these results could involve the fact that initial moistures in RR seeds may be slightly higher than conventional, so discrimination models were mainly driven from moisture differences. This affirmation, although could be logical based on this single study, cannot explain the discrimination among conventional and RR seeds in a previous study [10] involving a large population of seeds from hundreds of samples. Furthermore, when seeds have similar high moistures they are still differentiable. Figure 8 shows the PCA score plot (PC2 vs PC3) of the M97-304 variety scanned with the light tube, with both RR (pink, circled) and conventional (blue) seeds used for validation (similar moisture ranges involved). RR samples seem to have higher scores on the second PC. This pattern can also be seen, although not so clear, in the score plot from the Buchi FT-NIR instrument (figure 9).

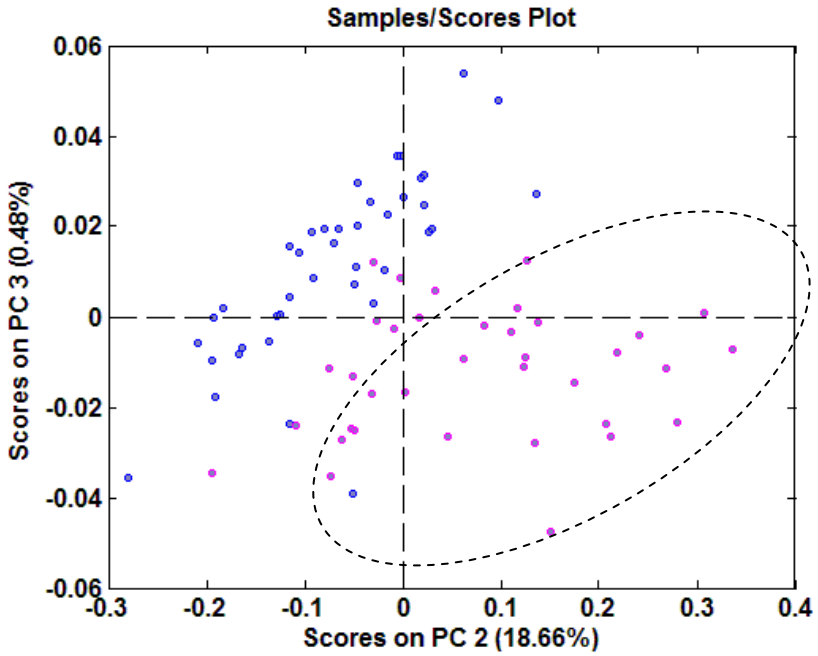


Figure 8. Score plot of seeds from the M97-304 validation set. RR seed spectra in pink and conventional in blue. Most of RR spectra are located in the circle.

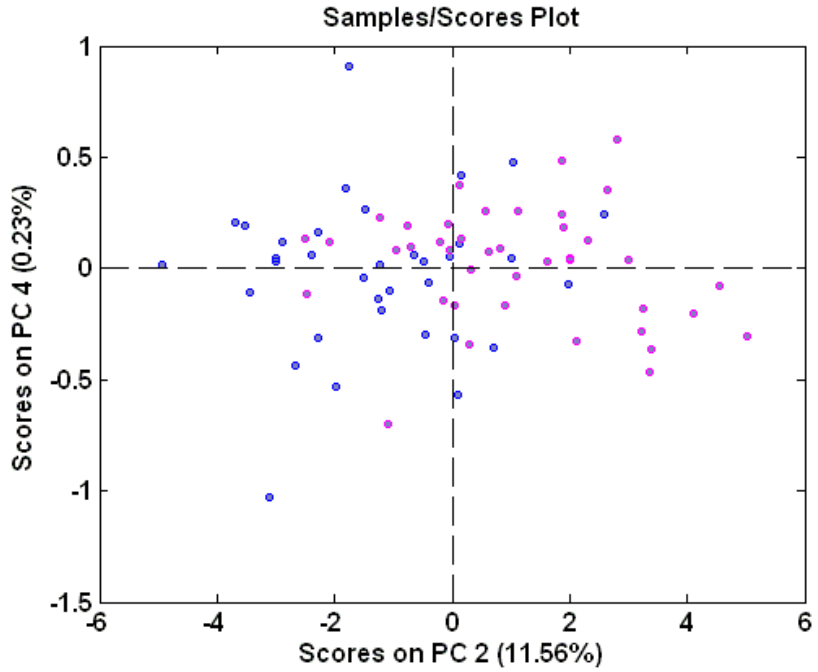


Figure 9. Score plot of seeds from the M97-304 validation set. RR seed spectra in pink and conventional in blue.

Another explanation is that there may be an interaction between water and fiber absorptions (carbohydrate) as they both happen in the wavelength range of 1400 – 1450 nm in the first overtone. In order to test this hypothesis, we selected the variety M97-304 spectra, both conventional and RR, and seeds from low and high moistures. We preprocessed the spectra with standard normal variate (SNV) in order to correct for offset and we averaged them, obtaining two spectra for sample M97-304 RR as a result of averaging dry and high moisture seed spectra, and similarly two spectra more for sample M97-304 conventional. The approach was as follows: The dry seed spectrum was removed from the high moisture spectrum. In absence of interactions, the water peak at 1,400 – 1,450 nm should be the only noticeable. Figure 10 shows the difference spectra for M97-304 RR (left plot) and M97-304 conventional (right plot). The absolute values of the difference spectrum are higher for the conventional sample because the wet conventional seeds had higher average moisture (17.8%) compared to the RR wet seeds (14.7%). However, it is interesting to see that in both cases, besides obtaining a peak in the carbohydrate absorption region which also includes water absorption (1,350 – 1,450 nm) another peak in the purely carbohydrate absorption region on the first overtone (1,150 – 1,250 nm) arose. The higher the moisture, the more relevant this last peak becomes, in correlation with the water peak. From this result we can conclude that the hypothesis of interaction between fiber and moisture eaxists and it affects the RR-conventional discrimination accuracies. Higher moistures may favor the identification of RR seeds decreasing the accuracy on conventional seeds.

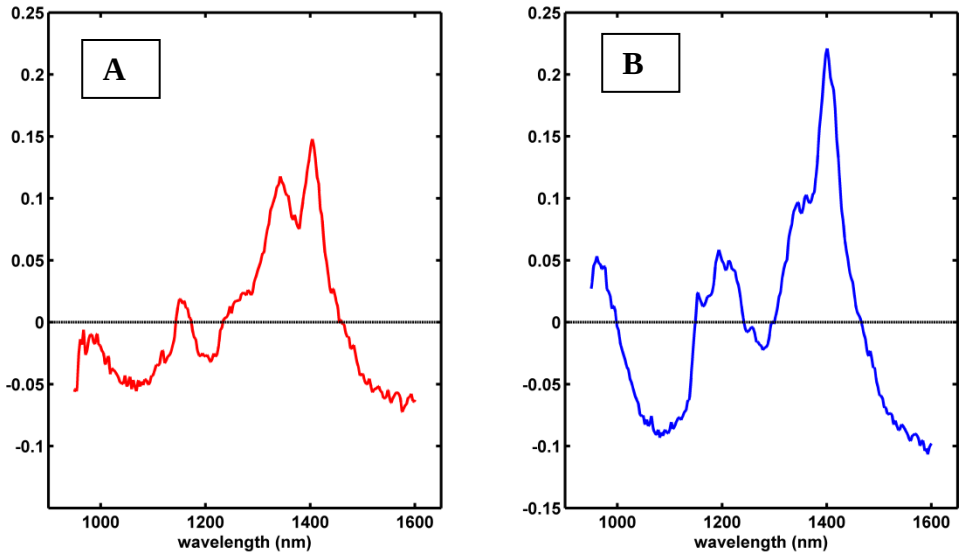


Figure 10. Difference between the average of wet seed SNV preprocessed spectra (150 seeds) and the average SNV preprocessed spectra of dry seeds (150 spectra) for sample M97-304 RR (left, plot A) and M97-304 conventional (right, plot B). Spectra was obtained from the light tube instrument.

CONCLUSIONS

The discrimination of RR and conventional dry seeds (moisture below 9%) using 5 varieties with their conventional and RR genetically modified versions and represented by 100 seeds each, lead to similar results compared to previous studies when using LS-SVM algorithms.[9,10] The performance of LS-SVM models were close to the performance of PCA-ANN models (lower eighty percents). LW-PCR lead to higher discrimination accuracies (over 95% for the light tube instrument) than LS-SVM and the LW-PCR models developed in the previous study.[10] The algorithm did benefit of having fewer samples with a larger number of seeds represented in the training set, instead of large number of samples with few seeds from each in the model as in the first study. However, it should be taken in account that LW-PCR models were optimized on the validation set and some of the best combinations of PCs and neighbors may lead to an overfitting of the models to the validation set. FT-NIR transmittance measurements were highly affected

by sample position and it negatively impacted the classification accuracies (78% with LS-SVM, 88% with LW-PCR) when compared with the light tube (82% with LS-SVM, 99% with LW-PCR. Within variety discriminations (same variety, conventional and with RR resistance gene) gave high accuracies for most of the classes. LS-SVM correctly discriminated the 92% of the seeds on average in both instruments, and discriminations were close to 100% with LW-PCR. This indicates that NIRS could be used for discriminating conventional and RR soybean seeds developing models for single varieties. However, some varieties could be easily discriminated (above 95% with LS-SVM) than others (around 75% with LS-SVM). So the application may not be usable for all varieties. Further work could develop models including seeds from different crop seasons and fields and test any changes in accuracies.

Varieties showing the highest misclassification rate often showed the lowest bulk moisture when compared to the total data set. Samples with higher moistures seemed to be easily discriminated. RR seeds were the most negatively affected by lowest moistures. However, when the discrimination models trained with dry seeds were validated with higher moisture seeds all the conventional seeds were misclassified as RR while most of the RR seeds were correctly classified. Moisture difference alone does not drive the classification of RR and conventional soybean seeds. Previous studies suggested that fiber was playing an important role in NIRS discrimination of RR and conventional soybean seeds.[9] Because the water absorption region on the first overtone (1,400 – 1,450 nm) overlaps with the carbohydrate (fiber) region (1,350 – 1,450 nm), a possible interaction of water-fiber may affect the discrimination. We proved by subtracting the soybean dry average spectrum to the high moisture spectrum that carbohydrate peaks in both the overlapping absorption region (1,350 – 1,450 nm) and the region exclusively involving carbohydrate absorption (1,150 – 1,250 nm) remain relevant. Changes in bonds and vibrations of carbohydrate molecules due to water absorption could have modified the absorption signal in those regions. The implication of this finding is the usability of discrimination models only at short moisture ranges. This is not a problem for breeders, who deal with seeds with similar moisture. The success of within variety

discrimination models with seeds with wide moisture range in the training set is unknown, as it is the algorithm that would perform the best in that situation.

REFERENCES

1. S. N. Cohen, A. C. Y. Chang, H.W. Boyer and R.B.Helling, “Construction of biologically functional bacterial plasmids in vitro”, Proc. Natl. Acad. Sci. 70, 3240 (1973)

2. A. Bakshi, “Potential adverse health effects of genetically modified crops”, J. Toxicol. Environ. Health B Crit. Rev. 6, 211 (2003)

3. C. Benbrook, “Impacts of genetically engineered crops on pesticide use: The first thirteen years”, *The organic center critical issue report* (2009)

4. R. Schenepf ,“Genetically engineered soybeans: acceptance and intellectual property rights, issues in South America”, *CRS report, library of the congress* (2003)

5. S. Konduru, J. Kruse, and N. Kalaitzandonakes, “The Global Economic Impacts of Roundup Ready Soybeans”, in *Plant Genetic/ Genomics, Genetics and Genomics of Soybeans*, G. Staycey, Ed. (Springer, New York, NY), vol. 2, part IV, p.375.

6. E. Folmer, J. Pedersen, and A. Holst-Jensen. GMO analysis methods in Nordic Council of Ministers (2004) Control of GMO content in seed and feed-possibilities and limitations. TemaNord 2004:541. Ekspressen Tryk & Kopicenter print.

7. J. Miljuš-Djukić, B. Banović, Z. Jovanović, D. Majić, M. Milisavljević, J. Samardžić, and G. Timotijević, “Abundance of soybean roundup ready modification in food and feed samples from Serbian retail markets”,Romanian Biological Lett**. 15(1)**, 102 (2010).

8. P. Hübner, H.-U. Waiblinger, K. Pietsch, and P. Brodmann, “Validation of PCR methods for the quantification of genetically modified plants in food”, J. AOAC Int. **84 (6)**, 1855 (2001).

9. S. A. Roussel, C. L. Hardy, C. R. Hurburgh Jr., and G. R. Rippke, “Detection of Roundup Ready™ soybeans by near-infrared spectroscopy”, Appl. Spectrosc. **55(10)**, 1425 (2001).

10. L. Esteve Agelet, A. A. Gowen, C. R. Hurburgh, and C. O’Donell, “Near Infrared Reflectance Spectroscopy applied to discrimination of conventional and roundup ready soybeans”, submitted to Appl. Spectrosc.

11. P. R. Armstrong, “Rapid single-kernel NIR measurement of grain and oil-seed attributes”, Appl. Eng. Agric. **22**, 767 (2006).

12. J. G. Tallada, N. Palacios-Rojas, and P. R. Armstrong, “Prediction of maize seed attributes using a rapid single kernel near infrared instrument”, J. Cereal Sci. **50**, 381 (2009).

13. L. Lukas, B., Hamers, B., De Moor, and J., Vandewalle, LS-SVMlab Toolbox version 1.5 (2003).

CHAPTER 4. FEASIBILITY OF NEAR INFRARED SPECTROSCOPY FOR ANALYZING CORN KERNEL DAMAGE AND VIABILITY OF SOYBEAN AND CORN KERNELS

A paper to be submitted to the Journal of Cereal Science

[1] Lidia Esteve Agelet, [2] David Ellis, [3]Susan Duvick, [4]Susana Goggi, [1]Charles R Hurburgh[*] and [3]Candice Gardner

[1]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames IA 50011

[2] National Center for Genetic Resources Preservation, Fort Collins CO 80521

[3] Department of Agronomy, Iowa State University, Ames IA 50011

[4] Seed Science Center - Experiment Station, Iowa State University, Ames IA 50011

***** To whom correspondence should be addressed at 1547 Food Science building, Iowa State University, Ames, IA 50011-1060. Phone: 515-294-8629, Fax: 515-294-6383, Email: tatry@iastate.edu

ABSTRACT

Current US corn grading system accounts for the percent of damaged kernels, which is carried by time-consuming visual inspection. Near infrared spectroscopy (NIRS), a non-destructive and fast analytical method, is tested as analytical tool for discriminating heat and frost-damaged corn kernels. Four classification algorithms were utilized: Partial least squares discriminant analysis (PLS-DA), soft independent modeling of class analogy (SIMCA), k-nearest neighbors (K-NN), and least-squares support vector machines (LS-SVM). The feasibility of NIRS for discriminating viable or germinating corn kernels and soybean seeds from abnormal or dead seeds was also tested. This application could be highly valuable for seed breeders and germoplasms because current viability test are

based on a destructive test of germination. Head-damaged corn kernels were best discriminated by PLS-DA, with accuracy equals to 99%. The discrimination of frost-damaged corn kernels was not possible; NIRS could not detect any difference between frost-damaged and sound kernels. Discrimination of non-viable seeds from viable was not possible either. Since previous results in literature are contradictory with the current results in damage discrimination, the seed damage extent in which NIRS can detect enough seed changes to carry out discrimination should be analyzed in the future. The viability discrimination results showed that in analyzing damaged seeds, NIRS is entirely discriminating based on changes in the seeds due to purely damage, without any correlation with the seed ability to germinate.

KEYWORKDS: corn kernel, near infrared, heat damage, frost damage, viability

1.INTRODUCTION

Corn (*Zea mays L.*) is the main feed for cattle in US (90% of total grains used for feeding) (USDA, 2010), and it is also processed in many end products for human consumption and industrial uses. Genetic traits, compositional characteristics, and overall quality of corn are factors which relevance varies depending on the final product and technological process. For instance, wet milling processes for ethanol production require of high starch corn; the quality of the prime material (properly dried, no cracks or broken kernels, absence of foreign material) is more important for wet milling than for commodity grain. Environmental conditions and other post-harvesting activities impact the final grain quality. High moisture, heal, freezing, and artificial drying are some of the factors that damage grains and negatively impact grain quality and economical value. Six corn grades in US establish a measurement of grain quality based on batch test weight and % total damaged kernels (USDA-GIPSA, 2001). For total damaged kernels, the maximum percentage of heat-damaged (by either excessive drying or not proper moisture adjustment during storage), broken, and foreign material (BCFM) in each grade is specified. The percentage of frost, sprout or mold damaged seeds is included in the total damaged seed category. Other characteristics which affect corn quality but are not

reflected in the US grading system are % waxy corn, stress cracks, and insect infestation. Although environmental conditions cannot be controlled and affect grain quality, adequate handling of grain during harvesting and storing help preserve it. Genetic manipulation impacts the quality. Seed breeders seek the genes of interest that led to the expression of a targeted physiological trait and carry out the improvement of future lines through genetic manipulation. Germplasms and seed banks are pools of plant genetic material which maintain genotypic diversity and provide genetic resources for breeders. Seeds are kept for several years with controlled storage conditions of moisture and temperature. However, seeds progressively age, losing the ability to germinate, and eventually die. This forces seed banks to periodically monitor seed viability in accessions. Regeneration of seed batches is needed when germination falls below 85% (Humeid et al., 1995).

Both quality control of grains and determination of germination of stored seeds are carried out by qualified personnel and are time consuming tasks. For instance, the official detection of damaged and infested kernels is carried out by visual inspection. The germination of seeds can only be determined by destructive tests which suppose the reduction of the number of stored seeds over time. Germination, in percentage of germinated seeds over the total tested seeds, is the estimated capability of a seed lot to produce normal plants with good vigor under favorable controlled conditions. Dormancy (%) is also referred as the percentage of hard seeds over the total tested which did not germinate during the germination test. This may be due to either intrinsic physical characteristics of the seed (seed coat and internal structures) or induced conditions by environmental changes. Both percent of dormant seeds and germination are required characteristics on certified seed package labels. The higher percentages the better, as it assures farmers higher productions in normal field conditions, and longer storage of the seeds. The germination test is the official method for testing seed viability, with several variants according to the conditions for which the seeds need to be tested for. The test also allows detecting abnormal, low vigor and dormant seeds. Tetrazolium dying of seeds (Cottrell, 1948) serves as a fast alternative for testing seed viability, but since it relies on the subjective evaluation of died seed structures it should be validated with germination

test (International Seed Testing Association, 1985). Some researchers have analyzed the germination process as an attempt to determining which seeds produce either abnormal plants or are dead in advance. Non-destructive study of changes in temperature profile during seed aging helped in the prediction of seed viability by Infrared thermography (Kranner et al., 2010). By iterative comparison of the thermal profile of a given seed with previously studied seed profiles, dead seeds were discriminated from alive at 85% accuracy, and heat-killed seeds from viable seeds at 100%. Water binding within seeds and on seed cells seems to be the most influencing factors in seed aging and deterioration (Becker, 1998), followed by processes of protein and lipid modification by oxidation (Bernal-Lugo and Leopold, 1998). Nuclear Magnetic Ressonance (NMR) was used to study the changes in tissue water and water-binding from viable and accelerated aging seeds (dead) (Krishnan et al., 2004), finding differences in the way that seed rehydratation happened but it concluded that more research and advances in instrumentation are needed in order to quantify the water status of intact seeds to determine seed viability status.

This paper analyzes the feasibility of Near Infrared Spectroscopy (NIRS) for discriminating heat and frost-damaged corn kernels from sound, and discriminate viable and non-viable corn and soybean seeds. Four algorithms were tested to find the most suitable discrimination method and understand the data characteristics: Partial least squares discriminant analysis (PLS-DA), soft independent modeling of class analogy (SIMCA), k-nearest neighbors (K-NN), and least-squares support vector machines (LS-SVM). NIRS is a technology which principle is based on the absorption of near infrared light by organic compounds and water. Some of the best well-known advantages of this technology are the high speed of analysis, low sample preparation requirements, and no destruction of the sample. Those advantages open the possibilities of applying NIRS for whole seed batch inspection and breeders purposes when used in single seed analysis. The limitations, on the other hand, are the initial dependence and reliability to an alternative external reference method (i.e. HPLC, GLC, Combustion…) and the high detection limits, which only allows NIRS quantifying compounds above trace concentrations. However, NIRS has been used in several applications for single seed

analysis with notable success. Quantitative applications in single corn kernel are mainly targeting oil (Orman and Schumann, 1992; Cogdill et al., 2004; Weinstock et al., 2006; Jiang et al. 2007; Janni et al., 2008; Spielbauer et al., 2009) but there are also researches developing calibrations of moisture (Finney and Norris,1978; Armstrong, 2006), starch (Spielbauer et al, 2009), and protein (Spielbauer et al, 2009). The predictive ability of those calibrations indicated that NIRS is a suitable technology for screening organic compounds in single corn kernels. The discriminative analyses of corn kernels found in the literature are based on endosperm characteristics (vitreosity, hardness). Good discrimination among kernels with vitreous and fliury endosperm have been reported (Campbell et al. 2000; Williams et al., 2009). Although a couple of studies tried the discrimination of kernels according to their toxin contamination setting different thresholds (Pearson et al., 2001; Dowell et al., 2002), those applications could only work at high accuracies discriminating sound kernels and kernels contaminated with toxin at high concentrations (>100 ppm for fumonisin, and >100 ppb for aflatoxin).

There are only a couple of studies which used NIRS for discriminating sound and damaged soybeans and wheat kernels. Wang et al. (2001) analyzed heat-damaged kernels using NIRS and could achieve very good accuracies (>95% of correctly classified) with just two wavelengths and partial least squares discriminant analysis (PLS-DA). They suggested that the classification was driven by differences in light scattering and color change in heat-damaged kernels. When carrying out a classification model for wheat based on vitreous and non-vitreous endosperms including defective kernels such as bleached, cracked and sprouted, bleached kernels were found to be the ones misclassified (Wang et al., 2002a). Wang et al. (2002b) also classified soybean seeds according the type of damage (sprout, heat, frost, mold or weather) with rates over 90% for most of the cases using artificial neural networks (ANN), but in that study heat damage classifications achieved lower accuracies compared to the ones achieved in heat-damaged wheat kernels (64%). Up to date, no publications have analyzed the use of NIRS for seed viability discrimination. Kusama et al (1997) used near infrared spectroscopy (NIRS) for analyzing ageing soybeans. They classified at 60% accuracy between sound and 3-day artificially- aged soybean seeds, 80% when aged for 5 days, and 100% when aged for 8

days. However, the correlation with seed viability was not considered. In this paper we analyze the possibility of detecting differences between viable (normal alive) and non-viable (dead and abnormal) seeds.

2.EXPERIMENTAL

2.1. Seed samples

2.1.1 Heat Damage

Twenty four corn kernels from nine accessions of nine different accessions (216 kernels total) were obtained and heat-treated in the National Center for Genetic Resources Preservation (Fort Collins, Colorado). The varieties were: NSL 2843, NSL 6528, PI 267209, PI 515179, NSL 32736, PI 167968, PI 213766, PI 176800, and PI 483549. Individual kernels were placed on a petri dish and microwaved for 45 seconds. Microwaving is a way to heat damage the kernels without obvious color change for most of the cases. Three of the seeds showed excessive heat damage and were not scanned. An addional set of 216 corn kernels from the same accessions were kept as sound seeds (not damaged).

2.1.2 Frost Damage

A hundred Frost damaged kernels from a single hybrid variety were obtained from the department of Agronomy in Iowa State University (Ames, Iowa). Fifty of those kernels were artificially frost damaged in their early growth stages (moisture content from 50 to 55%) when still in the husk. Ears were frozen in a Conviron growth chamber (Controlled Environment Limited, Winnipeg, Manitoba, Canada) in a 24 hour frost cycle. Damaged and sound corn kernels were stored in a cooler for several months at the same conditions before being taken out to achieve environmental temperature and being scanned.

2.1.3 Viability

The corn kernels and soybean seeds were obtained from the National Center for Genetic Resources Preservation (Fort Collins, Colorado). Three corn kernels accessions (three varieties: NSL 2837, NSL2838, NSL2842) and two soybean accessions (two varieties: PI79379 and PI132214) available in the bank were selected. Those seeds were stored at controlled temperature and moisture conditions since the early 1960s. Their selection was carried out according to the last viability results, dated from 2010 for corn samples and 2006 for soybean samples. Samples with similar percentage of germinated and no germinated seeds were desired for this study. The available germination results on record are shown in table 1.

Table 1.

Last available germination results for selected corn and soybean samples

Crop	Variety	Germination (%)
Corn	NSL2837	54.0
	NSL2838	56.0
	NSL2842	54.0
Soybean	PI79379	80.0
	PI132214	72.0

For each corn variety, 168 kernels were selected (total number of corn kernels analyzed equal to 504). Two hundred forty soybean seeds for each variety were selected, giving a total of 480 soybean seeds to be analyzed. All seeds were sent and scanned at the Grain Quality Laboratory at Iowa State University (Ames, Iowa). The kernels, individually identified in plastic plates of 24 wells for corn and 48 wells for soybeans, were sent back to the National Center for Genetic Resources Preservation (Fort Collins, Colorado) to conduct the warm standard germination test (7 days at 25C).

2.2. NIRS spectrometer and data collection

The instrument used for the three studies was a Perten DA 7200 (Perten Instruments, Inc., Springfield, IL). It is a diode array instrument which takes measurements from 850 nm to 1,650 nm, at 5 nm intervals (141 data points). The company provided a special single seed adapter consisting of a concave mirror surface (Fig.1) which can be inserted in the place of the regular bulk sample cup. The instrument was set to take three scans per each kernel after taking two blank readings, and provide the average spectrum. The seeds were placed with tweezers on the concave mirror with the germ facing up to the light and detectors.

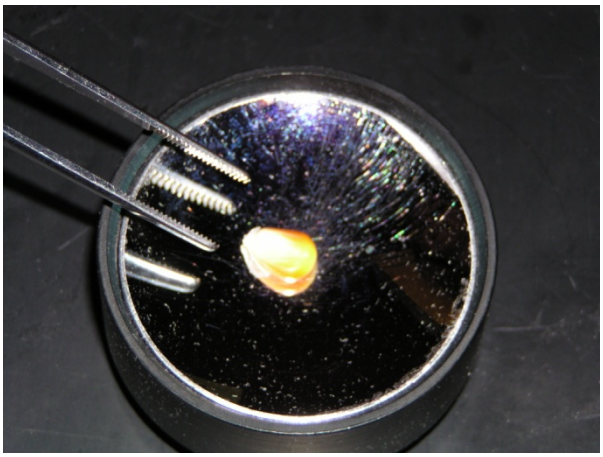


Fig.1. Perten DA 7200 single seed adapter. The corn kernel is located approximately on the center of a concave mirror using tweezers.

2.3. Data processing and discrimination models

Data from the instrument was imported with Jcamp format to The Unscrambler v.9.8 (Camo AS, Trondheim, Norway) for organization purposes. Matlab v.7.10 (Mathworks, Natick, MA) with PLS_toolbox v.5.8.3 functions (Eigenvector Research Inc., Wenatchee, WA) was utilized for data handling and developing the discrimination models. Two data

points from each side of the whole spectra were removed to reduce noise, leaving the working wavelength range from 860 to 1,640 nm. Possible outliers were visually detected by plotting the spectra and carrying out principal component analysis (PCA), where possible outliers would show high residual values and high Hotelli's T statistic values. For each of the studies, four classification algorithms following described were tested: Partial least squares discriminant analysis (PLS-DA), Soft Independent Modeling of Class Analogy (SIMCA), Least Squares Support Vector Machines (LS-SVM), and K-nearest neighbors (K-NN). The spectra was analyzed both raw (only mean-centering applied) and preprocessed with standard normal variate (SNV). This popular preprocessing method is known to reduce the scattering effects smoothing the noise on the signal, and has led to improvement in the results of several researches of single seed applications using NIRS and removed differences due to sample positioning (Weinstock et al., 2006).

2.3.1 Training and Validation Sets

For the first two studies of corn damage, the 75% of the spectra was utilized for developing the models and the remaining 25% (one spectrum picked every four spectra from the database) was kept for validation as test sets. The models were validated with the test sets and the accuracies were reported according the number of misclassified kernels in each study. Table 2 shows the number of samples for both training and validation in both studies.

Table 2

Final number of kernels (spectra) for each class utilized in each study. Class 1 are sound kernels, while class 2 are damaged kernels

Study	**Training Set**			**Validation Set**		
	Class 1	**Class 2**	**Total**	**Class 1**	**Class 2**	**Total**
Heat Damage	155	151	306	52	51	103
Frost Damage	37	38	75	13	12	25

The germination test results from the seed viability study are showed in Table 3; the percentage of dead seeds is low. The abnormal category makes reference to seeds which show damage or shredding in leaves, missing shoot or leaves, weak roots, or show impaired structures. Although those seeds are able to germinate, they show a lack of vigor which wouldn't allow long survival in regular environmental conditions. For this reason, those are considered non-viable together with dead seeds.

Table 3.

Germination results for corn kernels and soybean seeds per variety

Seed	Variety	Normal	Abnormal	Dead
Corn	NSL2837	111	37	20
	NSL2838	112	23	33
	NSL2842	109	28	31
Soybean	PI79379	190	41	9
	PI132214	159	46	35

Because some of the classification methods to be tested such as PLS-DA and LS-SVM can give biased results when the training class sizes are unbalanced (different number of samples in each class) and most of the seeds showed normal germination, we selected a reduced set of seeds from each variety for the training set in order to have similar number of samples per class (Table 4). The validation set included all samples left after creating the training set.

Table 4.

Final number of seeds (spectra) for each class utilized in each study. Class 1 are sound seeds, while class 2 are no viable. N stands for normal, A stands for abnormal, and D stands for dead.

Seed	Training Set			Validation Set		
	Class 1	Class 2	Total	Class 1	Class 2	Total
Corn	90 N	45 D 45 A	180	242 N	39 D 43 A	324
Soybean	50 N	25 D 26 A	101	299 N	19 D 61 A	379

2.3.2 PLS-DA Models

PLS-DA is a popular supervised classification method in NIRS applications because allows dealing with the highly correlated NIR variables (wavelengths). Similarly to partial least squares (PLS) for quantification, data reduction is conducted creating latent variables which are orthogonal with each other but at the same time trying to describe the response variable (in this case, the class labels). The SIMPLS algorithm offered by the PLS_toolbox v.5.8.3 functions was used (Eigenvector Research Inc., Wenatchee, WA). The class labeling was entered as a logical array, being each class represented by a column of zeros and ones (this last indicating the membership to one of the classes). The threshold for class separation was calculated according to the initial membership in each class. Through leave-one-out cross-validation, the optimal number of latent variables for the final model was selected looking at the fractional misclassification rate of each class and the root mean squared error of cross validation (RMSECV).

2.3.3 SIMCA Models

SIMCA works modeling each class independently by principal component analysis (PCA), a popular method known for reducing data dimensionality while keeping the

relevant information. For each class, an acceptance boundary defined by the maximum residuals of the samples from that group is created. Each class is expected to be modeled by a different number of principal components. We chose determined them by 10-fold cross-validation, being assessed by the eigenvalues, the Q and T2 values plot from cross-validation, and the Predicted Residual Error Sum of Squares (PRESS). When a new sample is presented to be classified, it is fitted in each model and the residual variance is calculated. If it is significantly higher than the average residuals for that class, the sample do not belong to that group. In this study the combination of Q and Hotelling's T^2 statistics are used for that purpose. The Q statistic is a measure of the residual between a sample in its initial dimension and its projection into the principal components. T^2 is calculated from the sample scores (projected samples) gives an idea of the variation of each sample within the PCA model. The sample will belong to the group where it will have a low Q and T^2. However, it is possible that a sample do not belong to any group or fits in both; in the first case, the PLS_toolbox algorithm assigns the class to the one which has the closest centroid in Euclidean distance. When the sample does not fit in any class, it is assigned to the group which leads smaller residual and T^2.

2.3.4 LS-SVM Models

Classification methods known as Support Vector Machines (SVM) were initially created for linear discrimination between two classes. The method is based on finding the widest margin of separation between classes. Thanks to the use of a mapping function (kernel function), this method can deal with complex classification problems which are not linear in the initial dimension but they may be at high dimensional spaces. We developed LS-SVM models using the LS_SVMlab v.1.5 toolbox functions for Matlab (Lukas et al., 2003). The Guassian Radial Basis Function (RBF) was selected as the non-linear mapping function. Two parameters had to be optimized: the kernel width (σ) and the model regularization parameter (γ), which is the trade-off between the margin width and tolerance to misclassification. The values of both parameters were simultaneously optimized using grid search ten-fold cross-validation. The input spectra were previously mean-centered, and the sample classes were defined by a binary vector of 1 and -1.

2.3.5 K-NN Models

For this algorithm, any new sample is classified according to the majority of vote from its "k" closest neighbors. The distance between two spectra, after mean-centering, is calculated from each wavelength as the difference in intensity between the two. The squared root of the sum of the squared distances at all wavelength points gives the Euclidean distance between the two spectra in equation 1, where the distance d between point i and j is calculated adding the differences in absorbance x between the two spectra at each wavelength l. The number of neighbors k is an odd number which we tested for 1, 3, 5, 7, and 9. The optimal number was selected by the one that lead to a lower leave-one-out cross-validation misclassification of the training set and then it was validated with the test set.

Equation 1.

$$d_{ij} = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2}$$

3.RESULTS AND DISCUSSION

3.1 Heat damage

The first thing to notice from heat damaged kernels was the overall low absorptions when compared with sound kernels (Fig. 2). This observation would agree with Wang et al. (2001) observation regarding the levels of light reflection and scattering between sound and heat-damaged wheat kernel being significantly different. However, offset and baseline differences can be easily removed by spectral preprocessing. In order to check for underlying differences, Savitzky-Golay second derivative (5 points gap and third order polynomial) was applied to remove baseline and overlapping effect (Fig. 3). The major absorbance differences between sound and heat-damaged kernels seem to arise in the carbohydrate regions, which could be easily explained by the fact that starch constitutes more than half of the corn kernel by weight and any change in this fraction may be more easily detectable than any change in the germ, which may also exist.

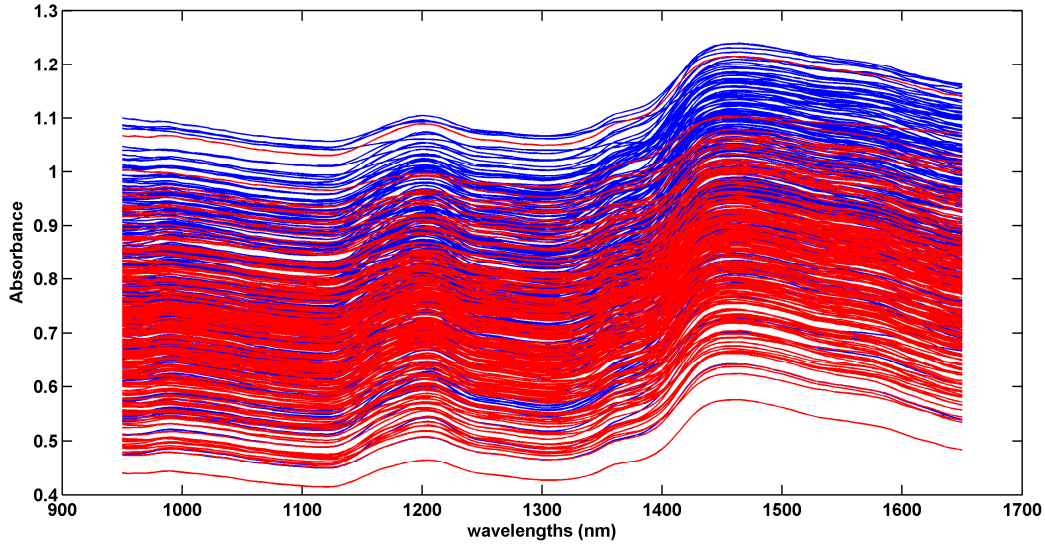


Fig.2. Heat-damaged corn kernels spectra (red) show lower absorptions than sound kernels (blue).

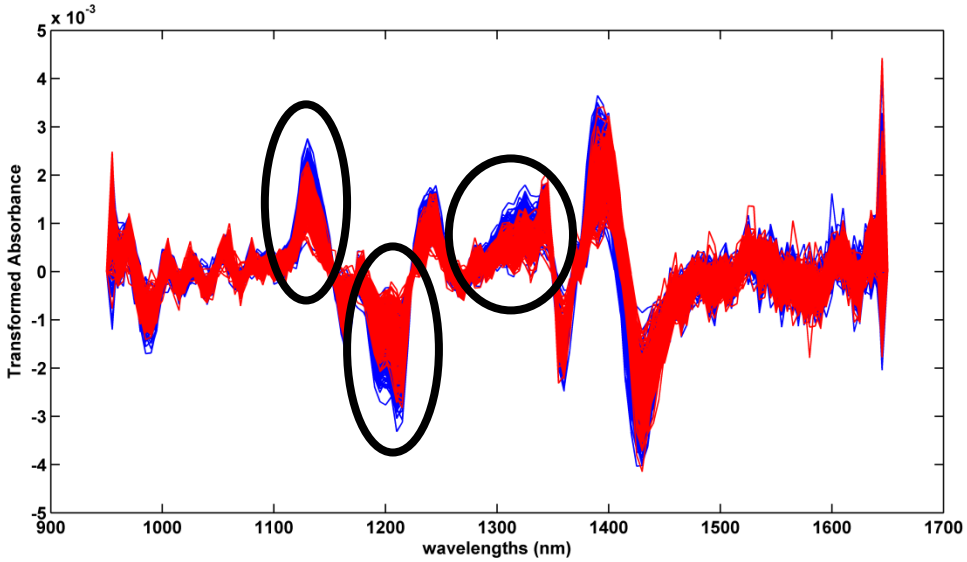


Fig.3. Savitzky-Golay second derivative of heat-damaged (red) and sound (blue) kernel spectra. The peaks where the major differences arise, which are in the wavelengths where carbohydrate absorption occurs, are circled.

Regarding the visual differences among spectra, the misclassification rates summarized in table 5 shows that not all the algorithms were equally successful in the discrimination. PLS-DA achieved the lowest misclassification rates with 8 latent factors (7 factors with SNV preprocessing), with almost perfect classification of the two classes. The regression vector plot shows also the relevance of the carbohydrate region in the classification model (Fig.4).

SIMCA achieved the second lowest misclassification rate when using raw spectra, using 9 PCs for heat-damaged kernels and 8 PCs for sound. When applying SNV preprocessing the misclassification rate rapidly increase. SNV may reduce the variability between classes due to light scattering and lead to similar PCA models for both classes. This again would agree with the relevance of light scattering in heat-damaged kernels classification (Wang et al., 2001). LS-SVM features for non-linear classification lead to higher misclassification rates but it was the only method which benefits of the use of SNV because the misclassification rate decreased considerably, indicating that it widened the lineal separation among classes in the projection hyperspace. K-NN models showed no difference when preprocessing with SNV, the algorithm performed the poorest. The best number of neighbors from cross-validation was initially spotted as 1, 3, and 7, with a total number of misclassified of 39. When tested on the validation set, 7 neighbors gave the best classification accuracies shown in table 5.

Table 5.

Misclassified corn kernels in the heat-damage study for each tested algorithm

Algorithm	**Raw Spectra**			**SNV Preprocessing**		
	Damaged	**Sound**	**Total (%)**	**Damaged**	**Sound**	**Total (%)**
PLS-DA	0/51	1/52	1.0	0/51	2/52	1.0
SIMCA	3/51	2/52	5.0	3/51	13/52	15.0
LS-SVM	6/51	9/52	15.0	2/51	3/52	5.0
K-NN	10/51	8/52	17.0	10/51	8/52	17.0

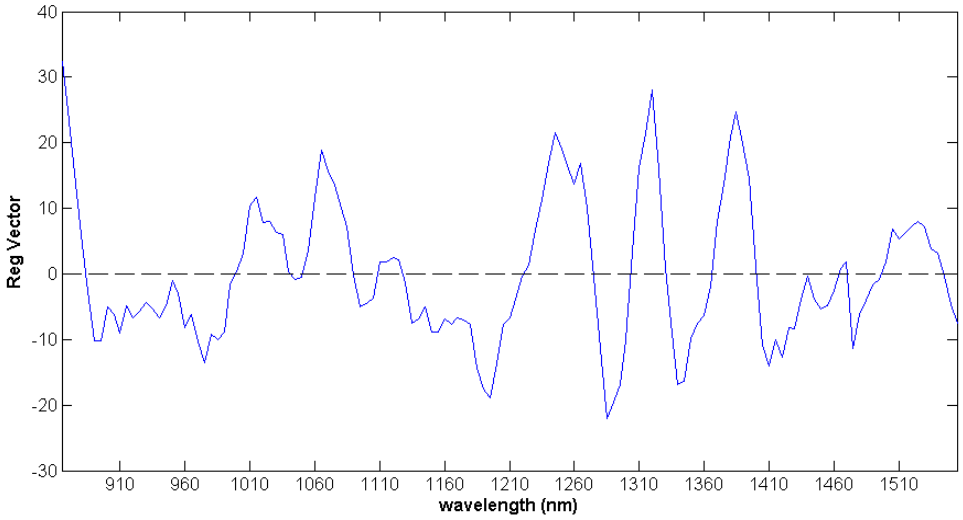


Fig.4. Regression coefficients (regression vector) from the PLS-DA model showing the largest absolute values in the carbohydrate region

Those results are better than the ones achieved by Wang et al. (2002) with soybean seeds and using artificial neural networks (ANN), who could only achieve classification accuracies on the upper 60%. On the other hand, the previous study by Wang et al. (2001) with wheat kernels lead to similar results (>97% accuracies by PLS-DA). They could get classification accuracies above 96% using just two wavelengths. Looking at figure 4 we could guess that using wavelengths from ranges of 1250 - 1400 nm could also lead to satisfactory classification accuracies, while best two-wavelength models from Wang et al. (2001) involved the wavelengths 985, 1,050, 1,550, and 1,575. Those wavelengths are mainly absorbed by protein. The differences in the relevant wavelengths in our regression model may be due to the heating method (hot air drying vs microwaving).

3.2 Frost damage

Differences among frost and sound corn kernels were not appreciable by the spectra and this was also shown in the discrimination results (table 6). The classification by NIRS was unsuccessful for all the discrimination methods, which lead to the same results. Sound and frost damaged kernels could not be differentiated and SNV lead to almost all the damaged kernels to be classified as sound for SIMCA and LS-SVM. PLS-DA optimal model used 5 latent variables with raw data and 3 latent variables with SNV preprocessing. Three and four PCs modeled frost and sound kernels respectively for SIMCA modeling. Both classes showed very similar PRESS and eigenvalues, which can be an indicator of the tight closeness of both classes. SNV preprocessing reduced the required number of PCs. (6 for frost damaged and 5 for sound), but the misclassification rate remained the same but bringing the two classes close, as all the misclassified kernels were damaged to sound. K-NN with both 5 and 7 neighbors gave the results shown in table 6 (K=7 when preprocessing with SNV).

Table 6.

Misclassified corn kernels in the frost-damage study for each tested algorithm

Algorithm	**Raw Spectra**			**SNV Preprocessing**		
	Damaged	**Sound**	**Total (%)**	**Damaged**	**Sound**	**Total (%)**
PLS-DA	6/12	4/13	40.0	6/12	2/13	32.0
SIMCA	6/12	4/13	40.0	10/12	0/13	40.0
LS-SVM	6/12	4/13	40.0	11/12	0/13	44.0
K-NN	5/12	5/13	40.0	3/12	7/13	40.0

3.3 Viability

The discrimination according viability was not successful for either corn or soybeans, and algorithm used (tables 7 and 8). The elevated misclassification rates show that discriminations were more random and no difference was detected by NIRS. Both PLS-DA and SIMCA models required again 4 latent variables or PCs. The use of SNV allowed using 1 latent variable more for PLS-DA and up to 7 and 8 PCs for SIMCA models but without enough improvement to lead to a successful application. Preprocessed spectra brought the two classes together for SIMCA models, similarly to the case of frost-damaged spectra. Reported K-NN results were achieved using a single neighbor.

Table 7.

Misclassified corn kernels in the viability study for each tested algorithm

Algorithm	**Raw Spectra**			**SNV Preprocessing**		
	Damaged	**Sound**	**Total (%)**	**Damaged**	**Sound**	**Total (%)**
PLS-DA	125/242	30/82	47.8	80/242	44/82	38.3
SIMCA	112/242	37/82	46.0	126/242	32/82	48.8
LS-SVM	133/242	28/82	49.7	124/242	43/82	51.5
K-NN	116/242	37/82	48.5	119/242	42/82	49.7

Table 8.

Misclassified soybean seeds in the viability study for each tested algorithm

Algorithm	**Raw Spectra**			**SNV Preprocessing**		
	Damaged	**Sound**	**Total (%)**	**Damaged**	**Sound**	**Total (%)**
PLS-DA	150/299	32/80	48.0	142/299	37/80	47.2
SIMCA	157/299	29/80	49.1	247/299	16/80	67.8
LS-SVM	166/299	37/80	53.6	133/299	35/80	44.3
K-NN	178/299	30/80	54.9	120/299	44/80	43.3

4.CONCLUSIONS

Among the three tested applications, only the discrimination of heat-damaged corn kernels was feasible. The achieved results are similar to the ones achieved by Wang et al. (2001) for wheat kernels, although in our study the carbohydrate region is the most relevant for the classification, probably for the difference in methods used for heat damage. PLS-DA performed the best, and more complex methods such as LS-SVM with RBF kernel mapping performed worse. SNV preprocessing, which has been useful for quantitative applications, did not lead to better results overall. It only benefit the LS-SVM algorithm, increasing the separation of classes in the higher dimension space where the data is mapped by the RBF kernel. K-NN, although performing the worse, gave discrimination accuracies very close to LS-SVM. Although the method is rarely used in NIRS studies, may have potential to be utilized in applications because of the convenience of gaving to optimize a single parameter (the number of neighbors).

Discrimination of frost-damaged corn kernels was not possible, even with the use of non-linear methods such as LS-SVM. Frost-damaged soybean discrimination by NIRS has been reported to be successful (accuracy over 90%) using ANN (Wang et al., 2002b). This arises the question of how sensitive is NIRS to detect damage in single seed or how appreciable the damage must be in order to be detected. Because high damage may affect seed viability, the test of viability discrimination served also to test the hypotheses that NIRS accuracy in discriminating damaged seeds depend on seed viability (the higher the damage, the higher the possibilities the seed is not viable). No differences were detected between sound corn and soybean seeds (viable) and naturally dead or abnormal (non-viable) with none of the tested algorithms. This indicates that NIRS can discriminate the considerable damage in seeds in terms of physical and chemical changes induced by the damage, but cannot detect changes merely caused by the dead of the seed (i.e. changes in water binding, oxidation of lipids, changes in protein). Seed aging has been proved to be tracked by NIRS (Kusama et al., 1997), but the threshold that separates aged and non-viable seeds is not differenciable by NIRS.

5.REFERENCES

Armstrong, P. R., 2006. Rapid single-kernel NIR measurement of grain and oil-seed attributes. Applied Engineering in Agriculture 22(5), 767-772.

Becker, H., 1998. Saving Seeds for the Long Term. Agricultural Research, 12-13.

Bernal-Lugo, I., and Leopold, A. C., 1998. The dynamics of seed mortality. Journal of experimental botany 49(326), 1455-1461.

B. Humeid, L. D. Robertson, J. Valkoun, and J. Konopka, 1995. Multiplication and rejuvenation of genetic resources at ICARDA. In: Engels, J.M.M., and R. Ramanatha Rao, R. (Eds.), Regeneration of seed crops and their wild relatives. International Plant Genetic Resources Institute, Rome, Italy.

Campbell, M. R., Sykes, J., and Glover, D. V., 2000. Classification of single- and double-mutant corn endosperm genotypes by near-infrared transmittance spectroscopy. Cereal chemistry 77(6), 774-778.

Cogdill, R. P., Hurburgh Jr., C. R., and Rippke, G. R., 2004. Single-kernel maize analysis by near-infrared hyperspectral imaging. Transactions of the ASABE 47(1), 311-320.

Orman, B. A., and Schumann Jr, R. A., 1992. Nondestrctive single-kernel oil determination of maize by near-infrared transmission spectroscopy. JAOCS 69(10),1036-1038.

Cottrell, H. J.,1948. Tetrazolium salt as a seed germination indicator. Annals of applied biology 35(1), 123-131.

Dowell, F. E., Pearson, T. C., Maghirang, E. B., Xie, F., and Wicklow, D. T., 2002. Reflectance and transmittance spectroscopy applied to detecting fumonisin in single corn kernels infected with Fusarium verticillioides. Cereal Chemistry 92(2), 222-226. Finney, E. E., and Norris, K. H., 1978. Determination of moisture

in corn kernels by near-infrared transmittance measurements. Transactions of the American Society of Agricultural Engineers 21, 581-584.

International Seed Testing Association, 1985. International rules for seed testing. Seed science and technology 13, 300-520.

Janni, J., Weinstock, B. A., Hagen, L., and Wright, S., 2008. Novel near-infrared sampling apparatus for single kernel analysis of oil content in maize. Applied Spectroscopy 62(4), 423-426.

Kranner, I., Kastberger, G., Hartbauer, M., and Pritchard, H. W., 2010. Noninvasive diagnosis of seed viability using infrared thermography. Proceedings of the national academy of sciences 107(8), 3912-3917.

Krishnan, P., Joshi, D. K., Nagarajan, S., and Moharir, A. V. , 2004. Characterization of germinating and non-viable soybean sedes by nuclear magnetic resonance (NMR) spectroscopy. Seed science research 14, 355-362.

Kusama, T. , Abe, H., Kawano, S., Iwamoto, M., 1997. Classification of normal and aged soybean seeds by Discriminant Analysis using principal component scores of near infrared spectra. Nippon Shokuhin Kogyo Gakkai-Shi 44(8), 569-578.

Lukas, L., Hamers, B., De Moor, B., Vandewalle, J., 2003. LS-SVMlab Toolbox version 1.5.

Pearson, T. C., Wicklow, D. T., Maghirang, E. B., Xie, F., and Dowell, F. E., 2001. Detecting aflatoxin in single corn kernels by transmittance and reflectance spectroscopy. Transactions of the American Society of Agricultural Engineers 44(5), 1247-1254.

Spielbauer, G., Amstrong, P., Baier, J. W., Allen, W. B., Richradson, K., Shen, B., and Settles, M., 2009. High-throughput near –infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. Cereal chemistry 86(5), 556-564.

United States Department of Agriculture (USDA)- Grain Inspection, Packers, and Stockyards Administration (GIPSA),2001. U.S. Corn inspection. Document

retrieved January 2011 from: http://www.gipsa.usda.gov/GIPSA/documents/GIPSA_Documents/corninspection.pdf

United States Department of Agriculture (USDA), 2010. Website information retrieved January 2011 from: http://www.ers.usda.gov/Briefing/Corn/

Wang, D., Dowell, F. E., and Chung, D. S., 2001. Assessment of heat-damaged wheat kernels using near-infrared spectroscopy. Presentation at the 2001 American Society of Agricultural Engineers annual international meeting, Sacramento, Ca, July 30-august 1.paper num. 01-6006.

Wang, D., Dowell, F. E., and Dempster, R., 2002a. Determining vitreous subclasses of hard red spring wheat using visible/near-infrared spectroscopy. Cereal chemistry 79(3), 418-422.

Wang, D., Ram, M. S., and Dowell, F. E., 2002b. Classification of damaged soybean seeds using near-infrared spectroscopy. Transactions of the American Society of Agriculture Engineers 45(6), 1943-1948.

Weinstock, B. A., Janni, J., Hagen, L., and Wright, S., 2006. Prediction of oil and oleic acid concentrations in individual corn (zea mays L.) kernels using near-infrared reflectance hyperspectral imaging and multivariate analysis. Applied Spectroscopy 60(1), 9-16.

Williams, P., Geladi, P., Fox, G., and Manley, M., 2009. Maize kernel hardness classification by near infrared (NIR) hyperspectral imaging and multivariate data analysis. Analytica chimica acta 653, 121-130.

CHAPTER 5. TRAINING ON NEAR INFRARED TECHNOLOGIES

A paper submitted to the Journal of Technology Studies

Lidia Esteve Agelet and Charles R. Hurburgh

Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011

* To whom correspondence should be addressed at 1547 Food Science Building, Iowa State University, Ames, IA 50011-1060. Phone: 515-294-8629, Fax: 515-294-6383, Email: tatry@iastate.edu

Abstract

Near Infrared Spectroscopy is a new technology that requires knowledge of several fields or disciplines. When training young students, there are some problems that we face. Inadequacy of current material for training and boosting the interest of learners are two of the most remarkable. We identify the drawbacks and limitations when training single students in this technology in our initial analysis and we expose an instruction system plan. Our approach counts on a hands-on activity or real-life problem, with the support of a training booklet created to cover student needs. Student mentors seem one of the keys of this system viability and efficiency, but the issue of training and preparing the mentors for their task still needs to be addressed.

Keywords: Near Infrared technologies, training, manual

Introduction

Near Infrared Spectroscopy (NIRS) is a relatively new technology based on the measurement of the absorption of near infrared light by organic compounds, which can be correlated with their concentration with the use of mathematical models. By developing a calibration model using reference data, which is basically a regression with more than one independent variable, this technology allows fast predictions of unknown concentrations of the calibrated organic compound in samples without the need of being destroyed or undergoing specific preparation for the analysis.

Although the discovering of the near infrared light region dates from the nineteenth century, it was not until the 1960s that it started being used as an analytical tool. The main reason for that delay was the lack of understanding of the chemical information held in that region, which was not directly interpretable and required advanced mathematical and statistical methods to be used for quantitative analysis. After the first successful applications for measuring moisture in seeds (Hart & Norris, 1962; Norris & Hart, 1965), there was an exponential growth of the use of the technology, which could be also attributed to the development of new algorithms and the fast improvement in computing power. NIRS is currently included in several analytical methods from the American Association of Cereal Chemists (AACC) and American Oil Chemical Society (AOCS) to determine moisture, protein, and oil in grains.

The Grain Quality Laboratory in Iowa State University has more than 20 years' experience with near infrared technology for grain analysis and other agriculture applications. It counts on a diverse group of instrumentation, big sample storage, and constant loads of new crop materials. Because it is currently well-known in the field, there are well-set internship programs and frequent visits of short term scholars interested in learning about near infrared technologies. The training activity in the laboratory is becoming one of the main priorities, but yet is challenging to combine the teaching task with the daily laboratory activities. Training personnel are involved in managing the laboratory or other research projects and this leads to low availability, thus an efficient training plan was long required. Firstly, our students needed some material to conduct

and support their learning experience while in the laboratory. Available material regarding theory and tips for successful analysis using NIRS is often very technical and oriented to professionals with a strong scientific background. As a result, our students became uninterested in the lecture, discouraged, and overwhelmed. Furthermore, inconsistency issues in the literature regarding terminologies increased the beginners' confusion. Secondly, monitoring of the learning achievements needed to be accomplished without requiring excessive attention from the lab personal, encouraging self-regulated learning. And finally, suitable projects and hands-on work had to be created according to the established expectations and final learning goals.

After struggling with a highly personalized, time consuming, and not fully efficient training system, we target the problems and critical aspects of training young students to a multidisciplinary technology such as NIRS with self-learning as an important component of the instruction. We identify confusing points of working with new analytical technologies which are still in process of being addressed by the scientific community. As a result of identifying those needs and bumps in the learning process, we suggest a training approach following an ASSURE approach (Analyze learners, set objectives, select material and media, utilize media, require learner participation, evaluate and revise) of instructional system design. As a prescriptive model, it provides a framework for organizing the process of creating instruction in a learner-centered, flexible, and dynamic manner. Although none of the current instruction system design models cited in the literature show a high efficiency and effectiveness as a whole (Blessing, 1995), the ASSURE approach seems the most appropriate under our circumstances (partial self-learning of a technology), as it gives special attention to the learners and media used. The final basic elements of our program were (1) a training manual for young students which boost a self-regulated learning about NIRS without the need of having previous background on the technology, (2) hands-on work in the laboratory and projects for new students to develop, and (3) a student's final presentation and report of the project to the laboratory personnel.

Analysis Phase: Approaching Young Students to NIRS and to Laboratory Dynamics

In the analysis phase of any instructional design model, the problems and needs must be identified. A comprehensive, detailed, and objective description of the needs is not possible and it always takes a high degree of subjectivity from the instructional designer (Lawson, 1980); In consequence, it is probable that needs change and new ones arise while the instruction is implanted. Similarly, proposed solutions may not be the best and may require an iterative process until the most suitable solution is found. Our initial analysis focused on several elements and characteristics summarized in Table 1, which were fundamental to setting both the final learning goals and using the right approach. Those are in accordance to Reigeluth's (1999) instructional situations, who stated that four main conditions drive the environment and outcomes of the instruction: nature of learners, nature of what needs to be learned, nature of learning environment, and nature of constraints.

Table 1. Relevant elements of analysis

Element	Key Characteristics
Learners	Age
	Nationality
	Background
	Previous knowledge
	Intrinsic characteristics
	Heterogeneity within group
Topic and Content	Multidisciplinary
	Complex
	Specific terminologies and concepts

Element	Key Characteristics
	Specific software
	Data handling
Environment	Staff
	Organization
	Resources
Training	Time
	Resources

Analysis of the learners: Young students and self-regulated learning

To know the learner characteristics before designing instruction is a relevant stage in ASSURE and the rest of instructional design models. In our case, that is extremely important because the resulting material should generate enough motivation and a stand-alone learning for the most part.

Our learners were targeted to be mainly in their early 20s, thus in their first years in college. Those are students very skilled in using new technologies such as computers and phones, yet it is reported that this new generation is lacking more ability of reflection (Prensky, 2001). Reflection is a key element in learning and adding new knowledge, helping students to relate new knowledge to prior understanding or choose strategies to novel tasks (Mezirow, 1991; Hmelo & Ferrari, 1997).

Encouraging and guiding learners towards a reflective process would benefit them not only in NIRS training but in their whole life as continuous learners. Values, habits and learning motivators among countries are known to be different (Howe and Strauss, 2000). In fact, Brookfield (1995) indicates that factors such as culture, political views, ethnicity and individual's personality have far more influence than the learner age inside the adult learner context. Narrowing the trainees' nationality is not feasible and it is not our intent

to do so because that would lead to discrimination. Motivation degree may be highly variable, as young students may still not have determined their career path and they would rather be interested in learning about the new environment and new country. In order to work on the motivation and enthusiasm of new students before they face any difficulty on the topic, the relevance and possible uses of the technology should be summarized and what they will achieve in their work is useful. Searching for a student's intrinsic interest and motivation may become a trainer's primary task, depending on the student, adjusting the extrinsic interest of any task according to the initial student motivation. Giving the student self-control of the activity without the responsibility being overwhelming or leaving them by themselves in a right measure requires close attention to each individual (Lepper, 1988).

The learner's background influences the way they will process the new information and how they will make a meaning of it (Leindhart, 1992). Learners who may already have some hints about the topic or may have contact to close disciplinary areas should still be checked in their current knowledge since there is a good chance there may be misconceptions that need to be addressed so they do not create new blocks of knowledge on the wrong basis (Dick, 1992). However, as we explain later, NIRS combines several fields and this imposed a limitation or additional problem for making the instruction the most effective even narrowing down the background of our trainees to agricultural engineering. Assuming all trainees will have similar background, heterogeneity due to programs and individual preferences will always remain. Setting assumptions regarding previous knowledge on any of the fields involved in NIRS technologies (chemistry, statistics, mathematics, computing…) were one of the old problems that haunted NIRS training although the flaw was mainly found in dumping too much information for such a short period of time and the final training goals. No previous information regarding the topic is assumed. Another factor related to the trainee is the stress to face. Students are often international and are hosted for a short period of time which would require adjustment to the new environment, language, and working system. Inevitably this impacts their attention span, already variable among different individuals.

Achieving Learner's preferences regarding delivery and environment, besides being

variable among individuals it was also much delimited for personnel's time constraints and laboratory environment, which cannot resemble conventional classes and cannot gather big student groups. A maximum of two or three students could be gathered at a time, but hosting a single individual was common. We had to rely on individual self-directed learning with external guidance for most of the training period. Self-learning and self-regulated learning are characteristics mastered in adult learning, but the disposition to this kind of learning involves cultural and personal factors still not well understood (Brookfield, 1995). It could be interpreted from the same author, that learners feel more comfortable with teachers from the same ethnicity, thus self-learning could avoid cultural shocks which would be otherwise observed in traditional learning.

Analysis of the topic: Dealing with a multidisciplinary technology

When dealing with Near Infrared Spectroscopy (NIRS), the term "chemometrics" is used frequently through all literature. The definition or meaning of this term refers to the use of computers, statistics, and chemistry. Basically, any sample containing organic compounds can be analyzed when it is irradiated with near infrared (NIR) light. Either reflected or transmitted light is measured (what is called spectrum), and it is either correlated through mathematical algorithms to the concentration of a specific compound (quantitative analysis and calibration of the instrument) or qualitatively analyzed for discrimination of samples according to specific characteristics. Without statistics and mathematics is impossible to use NIR light for analytical purposes, thus the introduction of equations and mathematical terms to novices becomes unavoidable. It is obvious that NIRS is a multidisciplinary field which requires some knowledge of at least basic chemistry, mathematics, statistics, instrumentation and informatics to become a successful professional in the field, develop any application, or simply carry out routine analysis. Furthermore, whenever the technology is applied to areas such as pharmaceutics, agriculture, or material science among others, specific knowledge from that field of work is required as well.

The threshold among disciplines becomes blurry as they all are part of the final body of knowledge and their interaction is a must for the growth of this field. This is a problem

according to the current education system as Wicklein and Schell (1995) related in their research. They set the need of reassembling topics and subjects such as mathematics, science and technology instead of considering them as segregated subjects and set clear boundaries. This fragmentation, according to Senge (1990), makes learners have a hard time connecting the pieces to create a whole concept and may be an impediment to achieving high-thinking solving problem skills, which require the use of different criteria and complex relationships (Resnick, 1987).

In conclusion, NIRS had to be introduced to our learners as a whole, as a topic that can only exists because of the fusion, correlation and understanding of several fields. Extreme care had to be taken to avoid excessive detail or relevance in any of the disciplines involved, especially when the topic is introduced for the first time and due to time constraints cannot be presented for the students to specialize and analyze in depth any of the aspects. This approach of equilibrated dose of each discipline in the topic is also optimal for the learner, who will have the chance to get a general taste and according to his/her interests and abilities will develop curiosity or stronger connections with certain areas.

We faced yet another problem: the adequacy of current material and information to be handed to the learners. We first used to provide our learners with additional reading material (papers, books, literature reviews), but we found that either those were not read at all or they were skimmed and no substantial learning was acquired. The language used in the literature seemed to be too far on the technical and scientific side for what our learners were ready to assimilate given their background. We think that one of the most frequent errors in current training materials, tutorials and reviews is trying to target a wide diversity of readers (or similarly, not target any specific reader at all). This approach works in scientific communities, where the driven professional seeks other sources to fill the knowledge gaps given their previous solid knowledge from a specific scientific area. This is not the case of young students, who are still building their knowledge bases. To round off, current training material deals excessively in depth with some or most of the areas (chemical theory, mathematics), which is neither unnecessary given the nature of our learners, who tend to reject any embellishment and considerably narrow down the

content to their learning safety zone (Blais, 1988). Another problem related to journal papers and reviews is the agreement in concepts and terminology, especially when they are retrieved from years of early NIRS stages. Because of the recent use of this technology and the growth of users and professionals from many different fields, the technology expanded in each disciplinary area independently for a while. Authors from diverse fields and even the same field were adopting their own terminologies and proceedings. One example involves statistics used for validation of the calibration models. Chemists and statisticians adopted, for a while, different terminologies for expressing the errors from the predictions, even if the meaning was ultimately the same. An agreement on what was the best way to report validation results (robustness and predictive ability of a calibration model) was not met among scientific papers either, leaving some applications and models incorrectly validated or not correctly reported to the scientific community. The use of acronyms and abbreviations has led to additional confusion. Most of the algorithms and methodologies in NIRS are composed of composite names that need to be abbreviated with acronyms, and while most become easily learned straight forward, a few have taken years to be unified in their spelling and meaning. For instance, this is the case of the abbreviation of near infrared as the light region (NIR) or near infrared technologies/near infrared spectroscopy (NIRS). It is also common to find papers where the measurement modes (either transmittance or reflectance) become part of the acronym of the technology (i.e. NIRT). Another case worth mentioning is a popular statistic in the NIRS community abbreviated RPD, which is used for describing the predictive ability of calibration models. Its original spelling by Williams (1987) was “the ratio of the standard error of prediction to the standard deviation” but because the spelling is so long and does not seem to connect with its acronym, other spellings such as “relative predictive determinant” can be found in the literature. Overall, different terminologies, acronyms and spellings can slow down the NIRS learning process and add confusion.

Another aspect to cover in NIRS training is the instrumentation. Spectrometers or spectrophotometers are easy to use in routine analysis if they have been previously calibrated. Instruments which have calibrations loaded are ready to provide fast

predictions right after the sample has been run (scanned). The process starts when pouring the sample to either a specific container or to the instrument hopper, entering the sample ID using a keyboard, and pressing a start button. The instrument automatically reads the sample absorbance, and through the load calibrations, displays the sample predictions (one from each calibration in use, usually one calibration per compound to be measured). Although the task is straight forward, students may encounter some problems associated with instrument conformation, sample characteristics, and software. Dealing with instrument warnings requires experience and knowing the instrument, which is a matter of time and a continuous process of overcoming difficulties, often requiring the instrument company support. There are some basic rules among instruments such as required warming up time, adequate sample presentation, and data base handling specifications.

Users working in routine analysis can leave instrument and calibration maintenance to external specialized companies since they are only interested in good measurements (predictions). No more data than the pop-out predictions is needed in the previous situation: a calibration model of interest has been previously developed and loaded in the suitable format for instrument recognition. Basically, the instrument will carry out sample readings or scans, will plug the data readings in the calibration model (which is basically an equation) and will display the prediction. Although routine analysis do not require more than that, NIR spectrometers can also store their readings at multiple wavelengths (spectrum). That data can later be used by the analyst to create calibrations or prediction models.

The process of exporting data is not very straight forward and each instrument stores the data in different formats. Furthermore, instruments support calibration models loaded in specific formats which can only be created using specific software and this leads to several software which are used to develop multivariate calibrations or regression models. Some are well known, such as SAS, R, or Matlab. Others (GRAMS, The Unscrambler) have been especially designed for spectroscopic applications. Instrument companies often have their own proprietary software, which pairs with the software used for data collection in the instrument, making the task of exporting data easier. Ultimately,

each instrument company encourages the use of their own software because they produce calibration models in a format directly compatible with their instrument. This would be no problem if all those proprietary software had the same logic and structure, which is not often the case. Furthermore, they tend to not be very flexible or intuitive to the users. The help command from software is an important yet forgotten tool when learning new software, but we have observed that for most software the help section is not very explanatory and fails to provide guidance or answer basic questions. For advanced applications, learning programming can be useful for processing and analyzing data. However, this is not strictly required since spectroscopy programs are evolving towards adding more algorithms without the need of programming. On the other hand, successful professionals in the field have been dealing with complex data avoiding the use of complex techniques and programming, concluding that for most of the cases a good understanding of the data and using the right combination of basic preprocessing methods is as powerful as using the most sophisticated calibration algorithms.

Analysis of the Context and Learning Environment

Context is the setting where acquired knowledge will be used or transferred into practice. We have already mentioned how in order to have a successful training leading the learners to reflection and context needs to be analyzed since it is a relevant factor when planning and designing instruction. Song et al. (2005) found that a reflective learning environment is one of the most important underlying factors behind development of reflective thinking in teenagers. Several years ago, Wittgenstein (1953) expressed that the best way to make learners understand and find the meaning of given information was to specify the final use of it, which is closely related to the final context. Later researches carried from a constructivist point of view of learning, agreed and target the contextualization of what has been taught as the key for motivation, understanding, and learning (Berryman, 1991; Bruner, 1966) and the lack of usefulness of whatever information which cannot be framed in any social, physical, or problem context (Dick et al., 2001).

It is difficult to predict future context of NIRS applications, as it can be used both in-situ

or rough environments (field, conveyors, fermenters) and laboratory or research-like activities. In our case, we can offer the learners the NIRS application for routine analysis of grain in a laboratory facility and the contact with on-going research projects. The laboratory is close in services and dynamics to any other laboratory in that field, although it has a more student-friendly environment since it is part of the university, and several full- and part-time students can be found working on a daily basis, both at graduate and undergraduate level.

One of the big pros is the chance of seeing and using a wide variety of instruments. Over 15 instruments with different conformations and from diverse companies can be found in the laboratory. The students have the chance to physically run the samples and become familiar with the laboratory dynamics and everyday activities. By immersion, students' learning is enhanced (Mourtos, 2003). Among those activities there are all the steps that new samples go through when they first reach the laboratory (identification, characterization, physical and compositional analysis, storage), meeting manufacturers and visitors from the grain industry, and attending staff meetings. The Learning environment, also known and learning context, is in both the laboratory and the office where the student will have an assigned desk and computer for his/her exclusive use. Other students, often carrying out different projects, are located in the same office so new scholars do not feel isolated. For an optimal learning environment, learners need to feel respected and safe in the environment (King, 2003). Although interactions are sought and encouraged through staff meetings and social activities, each student has to carry out a significant amount of self-learning.

In addition to the analysis of physical characteristics of the learning site (instrumentation, facility), the social aspect (student-oriented) and the relevance of the learning material to the hands-on work, Dick et al. (2001) suggest analyzing the support to the learners in their instructional system model. Learners of any age always need support, not only for assessment and to drive their learning to the right direction, but for motivation and encouragement (Rowntree, 1977). The laboratory manager and one or two more veteran students are assigned to support the new learner through the whole scholar term and respond to doubts and questions. They are introduced to the scholar project and carry out

a close follow up of the scholar activities and learning process.

Setting the Learning Goals

Once the analysis of relevant factors in the instruction have been done and needs and limitations have been detected, we had to figure out what we really wanted to teach or accomplish. We had to set some constraints and realistic learning targets, defining what our scholars should get out of the experience. Our overall final goal was for our students to understand how NIRS works and to develop their own calibrations for a given spectrometer data. We had to nail down the relevance of the material to what would allow students to answer how, with what, why and what can be predicted using NIRS. Some key questions and leading objectives are:

(1) What the technology limitations and advantages are.

(2) How to use basic instrumentation. The students should understand why sample characteristics are important and what effects may be caused when being scanned by the instruments.

(3) How to export data from selected instruments to specific software

(4) Which steps must be followed in developing calibrations.

(5) What basic algorithms/calibration models exist for calibrating.

(6) What basic methods can be used to mathematically process spectral data for signal enhancement or removing any signal noise.

(7) When those calibration models should or not be used.

(8) What is the best way to report any calibration performance.

Designing Material and Choosing Media

From the analysis thoughts and keeping in mind the relevance of the context, any instructional activity to be designed should resemble activities and experiences that learners may experience once they are done with the training for a more efficient transferring of skills and knowledge. Reflection, a high level thinking, is facilitated by contextualization but it is also associated with intrinsic motivation and positive attitudes.

While extrinsic motivation is driven by external forces (rewards, punishments…), intrinsic punishment is pretty much associated with the individual personality and willingness to learn. Both of them are equally important for learning (Lahey, 2007), but intrinsic motivation drives the individuals to work in more autonomy and be self-driven (Sheldon, 1995).Due to our constraints on personnel and the reduced number of scholars hosted at the same time, we had to enhance the intrinsic motivation of our learners making them enjoying the activities without expecting rewards, pass evaluations, be forced to compete and carry out tasks with a tight time frame, and have someone controlling them all the time. Those actions, according to Hennessy (2003), negatively impact intrinsic motivation. Following those rules, no reward or evaluation in a formal manner was created in our training approach but the informal encouragement and assessment to students at periodic basis, and the small assignment of sharing with the laboratory personnel what they accomplished in a presentation under an informal environment. Although the whole training period was defined, no time constraints were set for the student to carry out the activities but just the final presentation before they leave the laboratory. Finally, and as mentioned in the analysis phase, it is neither possible no feasible to have anyone surveilling the learners, although having enough support is equally necessary. For this reason, a couple of mentors or students were assigned to be there for any arising doubt and need of the new learner.

Although it is not so easy to target what is going to enhance the learning in each individual, it is certain that the designed activities and material should awake the student curiosity, not being excessively hard or too easy, have meaningful goals, and some degree of uncertainty (Lepper, 1988).

Developing a Training Manual

Which kind of material or technology is best for an instruction is not an easy choice and may not be the most relevant according to Bernard et al. (2004) and Tallent-Runnels et al. (2006) , who found that no differences between training delivery media (presencial, on-line, class-room) were found in terms of effectiveness, but what really matters is the learning environment and the instruction design (Clark, 1983). Again, laboratory

constraints in personnel availability influenced the way that information had to be delivered. In order to gain some time and get the students started, we developed a training manual which is sent to the students prior to their arrival to the laboratory. The main sections are summarized in table 2 and explained later with more detail.

Table 2. Main sections of the training booklet

Section	Learning Purpose
Near infrared story and theory	Terminology, origin, technology strengths and weaknesses
Instrumentation theory	Instrument conformations and suitability, measurement modes
Processing and calibration methods	Overview of algorithm steps, method suitability
Steps for calibration development	Procedures, stages, and tips for calibration development
Advanced topics	Additional, intrinsically motivated learning
Practical examples and exercises	Reinforce software use, dealing with diverse problems and data formats

Ideally, the first contact with the learning material should be self-directed, so when the student starts hands-on work he/she has basic knowledge to start building new one and draw connections between the blocks of knowledge. Often and unavoidably, this is not the case, and the first contact with the topic happens once the student starts in the lab. In any case, if the first contact happens by oral explanation of the laboratory manager of student mentors while having a laboratory tour, the assimilation is equally efficient since

the instrumentation and samples are physically present (contextualization).

The manual is intended to be a guide for the student through the training period and still be a guide after the training experience, for their personal use and consultation if they decide to continue on the field. It is designed to comprise the relevant aspects of theory and terminology, relevant sources of information for further research, a step by step guide for data analysis and calibration development, and practical exercises. It was relevant to constantly ask what was really relevant according to the established goals, and what could somehow be related to a practical aspect and application. If students fail in figure out the practical aspect of the exposed theory, they may not find the meaning of the material and consecutively fail in transferring it to the real context or in practical situations (Bransford, Sherwood, Hasselbring, Kinzer, and Williams, 1990).

We expect our young students to read the most detailed theoretical aspects by themselves and ask any possible questions or doubts, while focusing on the aspects they may want to learn more from. This would be in agreement with Dick's theory (1992), which gives the learners the freedom to select learning activities in favor of the theory of constructivism which supports the individual learner processing of information, and, as previously mentioned, this should be in favor of intrinsic motivation.

Near Infrared theory. We briefly introduce the story of near infrared as analytical tool and its evolution through the years. That brings the understanding that NIRS is a new technology which has experienced a big growth in the recent years due to the advances in statistical analysis, development of new algorithms, and increase in computer power. The chemical theory is introduced overviewing basic theory of atomic and molecular structure. Although knowing the chemical theory may seem not so useful in the beginning or not needed for practical applications, it is important to understand what a NIR spectrum is and how it looks like. That also helps recognizing each spectral region and the compounds involved in the absorption of certain light (protein, water…) and improve the process of calibration, targeting those spectral regions that seem more relevant to measure certain analyte. Finally, in this chapter we discuss different ways to measure near infrared absorption (by transmittance or reflectance). This provides a good

starting point for students to choose the best measurement mode when developing new applications and choosing instrumentation.

Instrumentation. This chapter introduces the basic instrument sections and up-to-date conformations. It is out of scope for our students to put hands on the internal section of the instrumentation in such early learning stages, but we consider useful to know the basic instrument conformations available in markets. Understanding the uniqueness of each instrument and advantages/weaknesses that each conformation brings to the analysis can make the students think about their suitability in developing further applications. Physic and optic set up behind the instrumentation is complex, and so they are individual parts such as lenses, detectors, gratings, or filters. There is a high variety of conformations, and there is not much updated material available which explain them. A good literature review from existent instrumentation and advances in the field was required to have a good up-to-date summary which students could consult. We intended to have our learners focusing on the instrument general conformation and data collection system rather than go too much in detail on the principles behind optics and electronic devices. It was also not considered appropriate to give instructions regarding the use of individual instruments because there are many instruments in the lab, some are temporal, new ones may arrive, and the one/s to be used will depend on the selected project. Furthermore, providing the instructions while performing the hands-on work task by the manager or mentor is much more efficient and it never gave problems. Since the use of the instrument is systematic, if there are any doubts during the process the learner can also consult other students in the lab and that helps them interacting.

Most of literature includes nice diagrams of the internal sections of the spectrometers, which are very useful in understanding the scanning and data collection process (how light irradiates a sample, is read back by the detectors after passing the optic set up, and finally received by the computer). One of the limitations of showing the diagrams is that their appearance is far from the real physical instrument, and students would not recognize the components when the instrument is open. Since there is not always the possibility to have a look on the internal section of the instruments, we gathered pictures

from each important section (manufacturer's websites or taken pictures) so students can relate the real piece with the diagram. This is another case of the previously mentioned contextualization: creating a strong learning connection between the real object (real world) after they understand its function in the instrument clearly motivate their learning (Mourtos, 2003). Figure 1 shows an example of an instrument conformation (pre-dispersive grating) with its relevant sections, and figure 2 shows the real picture of gratings. Having both diagram and picture allows students identifying the section whenever they have the chance to see the internal parts of an instrument in the future.

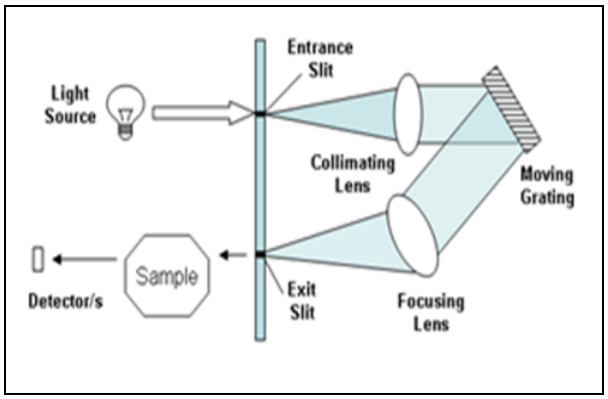


Figure 1. Diagram of a conventional pre-dispersive grating conformation of a NIRS instrument

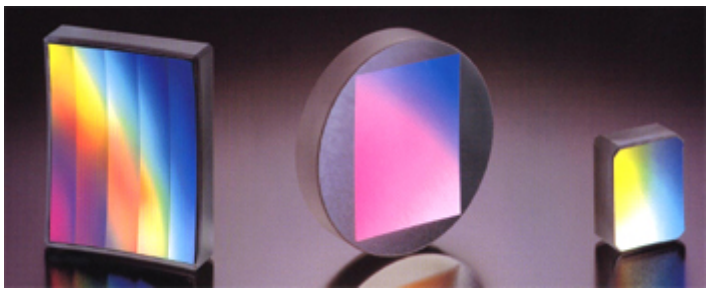


Figure 2. Picture of real diffractive gratings used in instruments (*Source: Hitachi High-Tech company web-site*)

*Overview of basic methods for calibration and data processing***.** We introduce the most common statistic methods and algorithms to carry out the data handling and analysis in developing calibrations. The redaction of this section has been done with care of using only strictly necessary equations. Our students often have an engineering background, and although their mathematical skills may be strong enough, the concept of multivariate statistics (working with highly dimensional data) differs from conventional modeling seen in engineering curriculum. The excessive use of equations is not advisable. According to Blais (1988), the conventional explanations of algebra with complex equations encourage an algorithmic activity far from the essence that learners perceive as important, so novices and new learners tend to select the minimum from the whole explanation that will allow them to achieve the mandated performance. This has been targeted as one of the reasons of the current mathematical atrophy. Summarizing, excessive in-depth explanations of both mathematical and chemical theories creates the follower role of students and a habit of dependence which does not let the student become expertise (Blais, 1988). Keeping the things basic and simple but yet induce to give some thought seem to be the right way to approach the algorithms and statistics of NIRS in our learners.

The visualization of the concepts by pictures and drawings is desirable for complex concepts as proven by Aso (2001). Figures and plots help explaining the effects on data of processing methods, for instance. We give some basic tips on how to choose the right algorithms and preprocessing methods, and point general differences in their impact to spectral data.

Steps for developing calibrations. The basic steps for developing robust calibrations are summarized in this section together with the introduction of popular statistics used to express the quality of any calibration model. Learners should be able to follow the given steps and advices in this section to create calibration models, and still be a useful guide for them in the future. In this section we introduce some aspects regarding (1) data formats and dealing with data cleaning, (2) enhance the relevant information and reduce

noise, (3) calibration process, and (4) validating and report the right statistics.

Advanced topics. This section gathers more complex preprocessing and calibration transfer methods (i.e. how to make data from two instruments be more alike, how to use the calibration from one instrument to another). Those are way beyond the learning objectives, but we still wanted to provide a hint of other kind of research for students who wanted to learn more – students acquiring high intrinsic motivation-.

Practical examples and codes. This section offers the student guided examples of the whole process of calibration procedure in connection to the previous chapter using different data types from different instruments. We provide the data to be used in several formats to be imported. We use two different software in the examples and exercises: The learners will start using The Unscrambler; Matlab examples are intended for advanced users who either have some previous experience programing in that language or want to use more advanced methods.

Hands-on project in the laboratory

We assign the scholars a project which represents either a current problem in the laboratory, testing a new instrument, or a new application of NIRS. The project is defined before the student starts the training, and the topic depends on the lab needs. In any case, the project involves hands-on work in the laboratory and use of instruments, data handling, and calibration development. Giving the students responsibility and a project that has a real use motivates them and through engaging them at emotional level, the learning experience happens faster and more efficiently (Leamnson, 2000). Furthermore, Bransford and Vye (1989) believe that learners have to use and experience what they have learnt by themselves. Not much variability among project difficulty is expected, although more or less success in the last results may be attained. This uncertainty in the final results, anyway, is what keeps the learner intrigued and motivated as previously indicated (Lepper, 1988).

Hands on work on each step of conventional sample analysis (from placing new

identification numbers to the samples, to finally scan them) underline the relevance of knowing the samples, being consistent, and follow certain operation order. This task not only helps students on not losing track of the data, but it is also required for the laboratory traceability system. Scanning the samples themselves and having to deal later with any mistake they did during sampling is a good experience that help them reinforce some working habits and important requirements when working with NIRS. Dealing with that extra work caused by mistakes and overcoming frustrations are important concomitants of the learning experience (Kort et al., 2001).

Dick et al. (2001) express how new skills and acquired knowledge is easily put in practice whenever there is support from managers and supervisors. Feeling supported is an important aspect for young students in order to maintain their initial level of engagement. This encouragement is part of the role of tutors. The current limitation for the task of assigning tutors and mentors is the individual personality of the available students and their formation or teaching experience. They are not especially trained and as students, may only get this task for a couple of years before they graduate and move on to their professional path. The best tutors are individuals that find challenging activities, point the problems and inconsistencies of student approaches to solve problems, and provide good analogies, examples or illustrations (Lepper , 1988). Excessive guidance does not make an individual a good tutor. Similarly to the negative effects of surveillance, excessive help and spoon-feeding to the student is not recommended by authors such as Weinert, Schrader and Helmke (1989), who point that learners who excessively rely on the mentors may not develop their self-regulating learning skills (i.e. how to plan and evaluate their own learning).

The sense of belonging to the lab group and interacting with other students and mentors has been also proved to enhance their performance and motivation. Not only by immersion as previously stated (Mourtos, 2003), but because higher achievements come from cooperative learning (Johnson & Johnson, 1989) because learners can process what they are learning from the theory and share their learning experience with other students (Johnson et al., 1998). Interaction with tutors and other learners brings better learning as students search for understanding all around Bowden (1990). This is one of our limiting

resources because as we mentioned, one or two students are involved in the training process. For this reason, tutors and their location in the office with other veteran students should allow them get cooperation. Relationships and cooperation with other working students in the laboratory reinforce the social nature of learning through keeping the general view of the activities against individual mastery (Leinhardt, 1992).

The routine use of instrumentation is easy, fast, and becomes a routine after several samples have been scanned but that period helps the students to get adapted with the whole environment and get used to the whole laboratory dynamics. After having all samples scanned by the selected instruments in the laboratory, we export the data for them. Although we briefly introduce possible data formats in the training manual, we often leave the tasks of exporting data from instruments as a voluntary task. The terminology used in the instrument software requires a good understanding of what the instrument set up is, what data you need to import, or which additional processed you want the instrument software perform on it (transpose, standardize, average…). Quite often, there are several exporting formats available to choose from. This has been one of the most controversial issues constantly reported by the NIRS community and which arise whenever using several instruments and software. Although text formats such as ascii or csv are very popular, there is a large list of proprietary data formats associated with the existent software. This makes the process of exporting data sometimes very confusing for new users, and undeniably inefficient for laboratories dealing with more than one instrument.

Data analysis with special software

Once the data is exported in the right format, the student is ready to work on their assigned computer in the office. The rest of their training period will be spent learning how to use the calibration software, organizing data, and finally developing the calibration/s. From the current software, we selected the one that could be easier to use, intuitive. The friendliness of the software must be taken in account. Friendly software has a spreadsheet based system with logic arrangement of options and uses standard terms. By similarity in the overall look and in the organization of option menus, it should be

similar to more popular programs such as Microsoft Excel, so users would get familiarized faster. In fact, students often start working with the data in excel, carrying out some sorting and cleaning tasks, and once they understand the data structure they proceed to export the data to the selected software. Excessive flexibility of software – characteristic that most powerful programs such as Matlab have - is not highly desirable for our new learners, who can become overwhelmed with too many options when choosing helpful plots and the analysis options during calibrations. At the end, our goal is to avoid the software to be a nuisance to overcome, but rather have a friendly structure which leads the user to only worry about understanding the calibration procedure.

The calibration development procedure was shown hands on once or twice by the tutor prior to having the training material, but we noticed how the amount of information we could give them through the process was not completely assimilated and thus the teaching was not very efficient. In order to be so, the process should be repeated a couple of times more which is not feasible. Students need to experience it by their own. Authors such as McKeachie's (1986) already reveal that in order to achieve active learning there must be active thinking of the student and not just sit and listen. Taking notes during the explanation do not let them pay enough attention to the practical examples and use of the software. The training manual was created to be a key material in this learning stage.

Despite software characteristics and the whole complex mathematic theory behind, students struggle with a feeling of uncertainty the very first time they start their own calibrations. The reason for that is the subjectivity on some of the choices during the process, for instance the deletion of bad samples or outliers. Some students do not feel confident to select samples that may be problematic. This issue is faced for more advanced users as well, when does a sample should be deleted? Sometimes it becomes a matter of trial-error and some hints and tricks are learned with experience. Making students understand that they can develop several models, resulting of trying several times with different processing methods, and understand that some of the choices are pretty much subjective takes several trials and encouragement. They ultimately should understand that what will specify the worthiness of the model is its validation. They should get to the point they can feel comfortable taking their own decisions whenever

they are justified. This phenomenon happens with other aspects during the calibration process. Having many choices and not having a right answer or rule to follow bring some anxiety on them. A final critical point is the validation procedure of the obtained calibration model. Their common doubts are regarding the use of the statistics, with the major question “how good my results are”. If they practice with grain samples, they can compare their validation statistics with several of the lab calibrations. Choosing the right statistics to report their validation and understand their meaning or significance has been shown to be one of the areas we have to put more attention to clarify.

Learner Participation and Overall Evaluation

Besides on-going assessment carried out by tutors and the manager, the overall evaluation of our training is done at the end, at the same time that the student presents in a staff meeting the project and the found solution where the student outcomes are analyzed, and compared with the outcomes expected from the instruction. Outcomes are, by definition, results that are measurable or observable. Most of the outcomes are shown on the go, while the student asks questions and expected reports, usually one or two. A close attention to real student outcomes give information regarding the efficiency of the training method: According to the instructional system design theory, learning outcomes should be evaluated. Evaluation of the learners through test of exams did not make much sense in this kind of instruction, but because learners should have some motivation and a small pressure to be on track, the laboratory checks the accomplishment of the goals making the learner present the results of the assigned project in an informal meeting. Knowing they have to present their results at the end, make them have a clear idea of the learning targets and create a small pressure or fear to disappoint, which positively impact their constant work according to Leamnson’s theory (2000). During the preparation of the presentation and get together of the results, our learners unconsciously carry out a self-assesment that allows them finding out where they stand and how far are from the goal. Together with formative assessment from the laboratory personnel, this process follows the one suggested by Chappuis (2005), who supports that students should be able to carry out their own assessment besides the one given by the mentors.

Conclusions

Teaching NIRS means dealing with a multidisciplinary field, and recognizing the extent of the theory to be thought is a critical point for young learners. Excessive theory is not useful, discourages their learning, and leads to confusion. Yet, not enough understanding may lead to incorrect use of the technology in the future, resulting in unstable calibrations, incorrectly reported and validated. We have summarized some of the critical aspects that make teaching this technology especially difficult with our reflexions and suggestions. Current papers and books have not been enough for our training needs so far, because they are often too technical for learners who are young and novice in the area. Furthermore, some inconsistencies can be found given the fact that the technology is relatively new. A user-friendly approach of summarized and practical material was needed so students could have some guidance to achieve the major learning goals mainly by self-learning given a hands-on task in our laboratory.

Our instruction is based on the assignment of a real-life project which starts with hands-on work in the laboratory and ends up with data analysis and a final exposition to the laboratory personnel. We created a booklet as a support material intended to be used through the whole learning process. More visual aids, fewer equations, use of more user-friendly language, reduction of theory depth, examples, and practical exercises using different data sets are some of the suggestions included in our booklet. One of the drawbacks is training young students individually due to time, personnel and economic resources. In this case, the role of assigning tutors or mentors is fundamental for training individual young students as they hold the key of encouragement and comfort in the laboratory. We still have to address the issue of instructing students assigned to be tutors.

References

Aso, K. (2001). Visual images as educational materials in mathematics. *Community College Journal of Research and Practice*, 25(5), 355 – 360.

Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A, Wozney, L., Wallet, P. A., Fiset, M., & Huang, B. (2004) How Does Distance Education Compare With

Classroom Instruction? A Meta-Analysis of the Empirical Literature. *Review of Educational Research*, 74(3), 379-439.

Berryman, S. (1991). Solutions. Washington, DC: National Council on Vocational Education.

Blais, D. M. (1988). Constructivism: A theoretical revolution in teaching. *Journal of Developmental Education*, 11(3), 2-7.

Blessing, L. T. M. (1995)**.** A process-based approach to design. IEE Colloquium on Wealth Creation from Design, 49:4, 1/4 – 4/4.

Bowden, J.A. (1990). Curriculum development for conceptual change learning: A phenomenographic pedagogy. (Occasional Paper No. 90.3). Melbourne: ERADU: RMIT.

Bransford, J., & Vye, N. (1989). Cognitive research and its implications for instruction. In L. Resnick and L. Klopfer (Eds.). Toward the thinking curriculum: Current cognitive research, (pp. 171-205). Alexandria, VA: Association for Supervision and Curriculum Development.

Bransford, J., Sherwood, R., Hasselbring, T., Kinzer, C., & Williams, S. (1990). *Anchored instruction: Why we need it and how technology can help.* In: D. Nix and R. Spiro (Eds.), Cognition, education, & multimedia: Exploring ideas in high technology (pp. 163-205). Hillsdale, NJ: Erlbaum.

Brookfield, S. (1995). *Adult learning: an overview*. In: A. Tuinjman (ed.), International encyclopedia of education. Oxford,UK: Pergamon Press.

Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.Chappuis, J. (2005). Helping students understand assessment. *Educational leadership*, 63(3), 39-43.

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, 43(4), 445-459.

Dick, W. (1992). An instructional designer's view of constructivism. In: T. M. Duffy and

D. H. Jonassen (Eds.), *Constructivism and the technology of instruction: A conversation* (pp. 17–34). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dick, W., Carey, L., & Carey, J. (2001). *The systematic design of instruction* (5th ed.) (pp. 88-114). New York, NY: Addison, Wesley and Longman.

Hart, J. R. & Norris, K. H. (1962). "Determination of the moisture content of seeds by near-infrared spectophotometry of their methanol extracts." *Cereal chemistry*, 39, 94-99.

Hennessey, B. (2003). The social psychology of creativity. *Scandinavian Journal of Educational Research*, 47(3), 253-271.

Hmelo, D., & Ferrai, M. (1997). The problem-based learning tutorial: Cultivating higher order thinking skills. *Journal for the Education of the Gifted, 20*(4), 401-422.

Howe, N. & Strauss, W. (2000). *Millennials Rising: The Next Generations*. New York:

Vintage Books.

Johnson, D.W. & Johnson, R.T. (1989). *Leading the cooperative school*. Edina, MN: Interaction.

Johnson, D. W., Johnson, R., & Smith, K. (1998). *Active learning: cooperation in the college classroom*. Edina MN: Interaction book company.

King, K. P. (2003). *Keeping pace with technology: educational technology that transforms. Vol 2: the challenge and promise for higher education faculty.* Cresskill, N. J.: Hampton Press.

Lahey, B. B. (2007). *Psychology: an introduction* (*9th ed.).* New York, NY: Mc Graw Hill.

Lawson, B. (1980). *How designers think (4th ed.).* Westfield, NJ: Eastview Editions.

Leamnson, R. (2000). Learning as biological brain change. *The magazine of higher learning*, 32(6), 34-40.

Leindhart , G. (1992). What research on learning tells us about teaching. *Educational leadership* 49(7), 20-25.

Lepper, M. R. (1988). Motivational considerations in the study of instruction. *Cognition and instruction*, 5(4), 289-309.

McKeachie, W. J., Pintrich, P. R., Lin, Y., & Smith, D. A. F (1986). *Teaching and learning in the college classroom: A review of literature*. Michigan: The University of Michigan.

Mezirow, J. (1991). *Transformative dimensions of adult learning*. San Francisco, CA: Jossey-Bass.

Mourtos, N. J. (2003). From learning to talk to learning engineering: Drawing connections across the disciplines. *World Transactions on Engineering and Technology Education*, 2(1), 1-7.

Norris, K. H. & Hart, J. R. (1965). Direct photospectrometric determination of moisture of grain and seeds. P*roceedings of 1963 International Symposium Humidity Moisture*, 4, 19 – 25. Reinhold, NY.

Prensky, M. (2001). Digital natives, digital inmigrants, part II: Do they really think differently. Retrieved January 4, 2011 from: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part2.pdf

Reigeluth, C. M. (1999) Instructional-Design Theories and Models: A New Paradigm of Instructional Theory, Vol. 2. New York, NY: Routledge publisher.

Resnick, L. (1987). Education and learning to think. Washington, DC: National Academy Press.

Rowntree, D. (1977) *Assessing Students: how shall we know them (2nd ed.)*.London, UK: Kogan page Ltd.

Senge, P. (1990). The fifth discipline: The art & practice of the learning organization. New York, NY: Doubleday/Currency.

Sheldon, K. M. (1995). Creativity and self-determination in personality. *Creativity Research Journal, 8*(1), 25-36.

Song, H., Koszalka, T. A., & Grabowski, B. L. (2005). Exploring Instructional Design Factors Prompting Reflective Thinking in Young Adolescents. Canadian Journal of Learning and Technology, 31(2). Retrieved January 4, 2011 from: http://www.cjlt.ca/index.php/cjlt/article/viewArticle/140/133

Tallent-Runnels, M. K., Thomas, J. A., Lan,W. Y., Cooper, S., Ahern, T. C., Shaw, S. M., & Liu, X. (2006) Teaching Courses Online: A Review of the Research. *Review of Educational Research*, 76(1), 93-135.

Weinert, F. E., Schrader, F. W., & Helmke, A. (1989). Quality of instruction and achievement outcomes. *International Journal of Educational Psychology*, 13(8), 895-912.

Wicklein, R. C., & Schell, J. W. (1995) Case studies of multidisciplinary approaches to integrating mathematics, science, and technology education. *Journal of technology education*, 6(2), 59-76.

Williams, P.C. (1987). Variables affecting near-infrared reflectance spectroscopy analysis. In P. Williams and K. Norris (Eds.), *Near-infrared Technology in the Agricultural and Food Industries*, (pp. 143–167). St. Paul, MI: American Association of Cereal Chemists, Inc.

Wittgenstein, L. (1953). Philosophical investigations. NY: Macmillan.

CHAPTER 6.GENERAL CONCLUSIONS

1. The feasibility of near infrared technologies to discriminate Roundup Ready® genetically modified soybean seeds from conventional was proven. Models developed with reflectance instruments and over 150 varieties from each class from several crop years and 15 seeds per class achieved accuracies between the upper 80% and lower 90% range when validated with seeds from samples represented in the training set.
2. There was no difference among classification algorithms when models are validated with seeds from varieties not included in the training test. Either locally weighted principal component regression (LW-PCR) or artificial neural networks (ANN) gave accuracies in the 70% range.
3. The USDA reflectance light tube instrument performed the best. When developing models with fewer varieties and represented by 100 seeds each, accuracies using LW-PCR algorithm were above 95%. However, extreme care must be taken as LW-PCR may be prone to overfit. Improvement of the optimization method is advised.
4. The low resolution imaging instrument and the transmittance instrument performed the worst. The imaging unit besides requiring more data manipulation and showing longer scanning times, it had the lower overall accuracies. The transmittance instrument showed high sensitivity to seed positioning, which negatively impacted discrimination accuracies.
5. The best technology should work on reflectance mode to avoid sampling variability affecting transmittance measurements, and take the spectra from the whole seed instead of just from one angle (e.g. conventional single point instruments).
6. In the second study with fewer varieties represented with more seeds, similar results to ANN models from the first study were achieved by modeling with support vector machines, with discrimination accuracies of 78% for the Fourier transform transmittance instrument (FT-NIR) and 82% for the USDA light tube.
7. Classifications within a same variety with and without the resistance gene gave accuracies above 90% for three of the varieties, with all instruments and both SVM

and LW-PCR models. However, two of the samples were not easily discriminated, with accuracies ranging 73 to 88% for both instrument and both discrimination models. The application of NIRS for Roundup Ready® and conventional varieties may be a useful tool for breeders, as best accuracies were achieved classifying seeds within a single variety. However, not all varieties may be discriminated with high accuracies.

8. High misclassification rates seemed to be correlated with low sample moistures. It was later proved that moisture affected classification accuracies, higher moistures leading to higher discrimination accuracies for Roundup Ready® seeds and higher misclassification rates for conventional seeds. We proved that there is a fiber-water interaction, and since fiber is a relevant compound in the discrimination, seed moisture is a relevant factor to control. The moisture of seeds to be tested should be within the training set moisture range.
9. Heat-damaged corn kernels could be discriminated at 99% accuracy by partial least squares discriminative analysis (PLS-DA), similar to previous results with wheat kernels. However, frost-damaged kernels could not be differentiated from sound. This result was opposite from the good classifications (above 90%) obtained in a previous study with soybean seeds.
10. The discrimination of viable (sound) and non-viable (abnormal and dead) was not possible. This indicates that NIRS damage discrimination does not involve seed viability, but only involves physical or chemical changes induced by damage. The extent of detectable damage by NIRS is unknown and should be analyzed in future researches.
11. We exposed the current training system and the training limitations when teaching NIRS in the grain quality laboratory. The most problematic aspects are the new terminologies used in the field, inconsistencies in terminologies, and null, incomplete or invalid reporting of model validation. When teaching young students and self-driven learning is needed, the role of mentors and hands-on activities is enhanced. However, mentors should receive proper training as well.

APPENDIX 1. LITERATURE REVIEW PAPER

Critical Reviews in Analytical Chemistry, 40:246–260, 2010
Copyright © Taylor and Francis Group, LLC
ISSN: 1040-8347 print / 1547-6510 online
DOI: 10.1080/10408347.2010.515468

A Tutorial on Near Infrared Spectroscopy and Its Calibration

Lidia Esteve Agelet and Charles R. Hurburgh, Jr.
Department of Agriculture and Biosystems Engineering, Iowa State University, Ames, Iowa, USA

Near infrared spectroscopy (NIRS) has had rapid usage growth since its first application in the 1960s in the grain industry. Since then, material science, food, environment, medicine, pharmaceutics, agriculture, archeology, and others have reported successful applications of near infrared technologies. Evolution and improvement of instrumentation is expected to continue with increasing users and applications. This review provides a guide for NIR beginners. Theory of NIR measurements, instrumentation, and stages of calibration development are covered in a non-mathematical approach, focusing on critical operating processes to provide new users a starting point for application and study.

Keywords NIR, spectrophotometer, chemometrics, calibration

INTRODUCTION TO NEAR INFRARED SPECTROSCOPY

Light is electromagnetic energy defined by the properties of wavelength, frequency, and energy. The energy content is indirectly proportional to the light wavelength—short wavelengths being more energetic—and is directly proportional to the wave frequency. Light can be arranged by any of those properties to form the electromagnetic spectrum. The infrared region is located in the middle of the electromagnetic spectrum, and has three major regions: far-infrared [300 GHz (1 mm) to 30 THz (10 μm)], mid-infrared (30 to 120 THz or 10 to 2.5 μm), and near-infrared (from 120 to 400 THz or 2,500 to 750 nm). The absorption by a material of any wavelength induces molecular vibrations. Changes in light energy between the three regions lead to varying absorptions from different molecules and bonds and induce different types of vibrations. For instance, the least energetic far infrared region light (FIR) is absorbed by heavy atoms, such as some inorganic and organometallic substances, while the mid infrared region (MIR) is popular for organic chemical analyses.

The near infrared region (NIR) is the most energetic infrared region and is close to the visible region in the electromagnetic spectrum, as discovered by Herschel in 1800. His experiments in measuring the heat produced by filtering the sun light on colors with a thermometer lead him to realize that temperature increased from going blue (450–475 nm) to red (620–750 nm). Temperature kept rising even after positioning the thermometer further from the visible red, which meant that more energy was present beyond the visible spectrum (1). Further significant research on the NIR region was not done for 150 years. MIR in analytical chemistry became popular, while the NIR region was ignored as it was considered to lack relevant chemical information: NIR spectra from any sample showed broad and overlapped low intensity bands, between 10 and 100 times attenuated compared to the sharper MIR fundamental absorptions (2). NIR broad peaks could not be directly assigned to specific chemical compounds or interpreted in a straight-forward manner as MIR spectra. Term spectrum (or its plural spectra) is commonly used in spectroscopy and it will be often used in this review to refer the light intensity measurements (after being either reflected or transmitted through a sample) as a function of wavelength.

NIR spectra are formed of overtones and combination bands. Overtones are electron excitations to higher energy levels which occur at multiples of the MIR fundamental frequencies. The entire NIR spectra contains up to four overtones (although the fourth overtone is very weak and ignored) from the absorptions of methyl C-H, aromatic C-H, methylene C-H, methoxy C-H, carbonyl associated C-H, N-H from primary and secondary amides, N-H from amides (primary, secondary, and tertiary), N-H of amine salts, O-H (alcohols and water), S-H, and C=O groups (3). Note that all those groups are found in organic molecules and water and absorb in MIR; hence, each NIR overtone repeats the chemical information of MIR but with absorption bands decreasing with overtone level. The combination bands region is located at higher NIR wavelengths (1900

Address correspondence to Lidia Esteve Agelet, 1545 Food Science Building, Iowa State University, Ames, IA 50014, USA. E-mail: lesteve@iastate.edu

to 2500 nm), and basically involves a combination of vibrations from the same chemical groups of the overtones, but as a result of interactions between molecular vibrational frequencies, overlapped information from Fermi resonances, and inactive MIR bounds among other phenomena (4). The bottom line is that chemical information in NIR spectra is repeated and highly overlapped through the whole wavelength range, a fact that discouraged researchers for a long time.

The beginning of the 1960s was an inflection point for NIR spectroscopy. Karl Norris, known to be the pioneer on NIR analytical development, and his U.S. Department of Agriculture team could determine moisture content from seed extracts (and later for whole seeds) using NIR bands with a multi-variate calibration approach (5, 6). In the early 1980s, the use of NIR by reflectance became an analytical, recognized method by the American Association of Cereal Chemists (AACC) for measuring protein in wheat, and the list was later expanded with methods for protein and oil determination in ground and whole soybeans, hardness determination of wheat, and protein content in small grains in general (7). The impact of Karl Norris' work was huge not only on the grain sector where NIR would prove to save significant time and money (8), but also in other non-agricultural fields such as pharmaceutics, polymers, material science, medicine, art, textiles, animal feed, and food where NIR is still leading to an extensive variety of applications and emerging NIR-related technologies.

This review intends to provide a summary on the basic principles behind NIR and its analytical use. Fourier transform NIR and NIR chemical imaging are briefly discussed as two of the most popular NIR-related technologies. Since the exponential growth of NIR applications can be mainly attributed to advances in instrumentation and data analysis methods, instrumentation and calibration development sections are emphasized using a practical, user-oriented approach. Critical calibration development stages are listed and common statistical methods discussed.

USING NIR LIGHT IN ANALYTICAL CHEMISTRY

When a sample is irradiated with light, according to energy conservation law, fractions are reflected, transmitted, and absorbed all summing to 1.0. The proportions depend on the light wavelength and sample properties (composition and thickness among others). Beer's law, well-known in molecular spectroscopy, defines the correlation of analyte concentration with its absorbance at specific wavelengths. Beer's law is not directly applicable in NIR spectroscopy because of several restrictive assumptions: no correlation between multiple absorbers, homogeneous samples, negligible light scattering, and constant path length. Notwithstanding this, Beer's law implication is still held by NIR analysis, but some further detailed challenges must be overcome first.

Although absorbed light cannot be directly measured, transmittance and diffuse reflectance can be correlated to light absorption according to Eqs. 1 and 2, respectively.

$$Apparent\ Absorbance = \log(P_o/P) = \log(100 / T(\%)) \quad [1]$$

$$\begin{aligned} Absorbance &= -\log(R_{relative}) = \log\left(\frac{1}{R_{relative}}\right) \\ &= \log\left(\frac{R_{standard}}{R_{sample}}\right) = \log(1/R_{sample}) \\ &\quad + \log(R_{standard}) \end{aligned} \quad [2]$$

Transmittance (T) is defined as the ratio of radiation passing a sample per unit area (P) divided by the initial radiation power (P/P_o), expressed as percentage. Log(1/T), also known as optical density, is called apparent absorbance because the effects from light dispersion in the sample are not taken in account. Although it is a close approximation, it is not exactly the same as the absolute absorbance as some of the emitted radiation is reflected before being transmitted (9).

The reflected light fraction shows higher complexity. There are two main components of reflected light: specular and diffuse. The specular component angle of reflection is the same as the incident light, is reflected to a single direction, and achieves its maximum intensity when the irradiated light is perpendicular to a smooth sample surface. It lacks NIR relevant information due to its minimum contact with the sample. The NIR diffuse reflectance component refers to the part of the incident beam that achieves a certain degree of sample penetration, it is scattered within the sample, and returned to the surface after within-sample absorption. It can be correlated to absorbance through Eq. 2. Relative reflectance ($R_{relative}$) is measured as the ratio of the sample measured reflectance (R_{sample}) over the measurement from a highly reflective material ($R_{standard}$, with reflectance approximately 100%) such as Teflon or Spectralon.

Transmittance measurements are best taken at lower wavelengths because they are more energetic, have more penetration power, and the absorption is weaker. Instrumentation that works in transmission mode works with a shorter wavelength range, usually not higher than 1800 nm. Measurements by diffuse reflectance are best taken at wavelengths between 1200 and 2500 nm. Above 2500 nm (MIR region), sample absorption becomes very strong. MIR measurements by transmittance cannot be carried out on thick samples; signal to noise ratio is reduced (10).

Summarizing, NIR measurements for analytical purposes can be carried out in two modes: transmission or diffuse reflection. Diffuse reflection mode allows working with thicker and denser samples without inducing as much heating as transmission. While sample path length is pre-determined and must be kept constant for transmittance measurements, the minimum sample required in reflectance mode is highly dependent on the wavelength range used in the analysis and sample characteristics such as density or packing, particle size, and material

absorption (11). Physical characteristics affect reflectance measurements especially at higher wavelengths (combination bands region); hence, any sample changes will create an additional source of variability and noise in the measurements (12).

Overall, reflectance measurements show a shorter dynamic range compared to transmittance (lower sensitivity) because information provided by diffuse reflectance originates from smaller sample portions and has been attenuated (13). Its repeatability is slightly worse which is more noticeable in heterogeneous samples. In specific applications, those limitations may not create significant errors, or may be mitigated by using of a wider range of wavelengths (14). Transmittance measurements exceed the accuracy of reflectance measurements in most pharmaceutical applications, although analytical sensitivity, signal to noise ratio, and limit of detection is highly affected by sample position and changes in geometry (15). Comparison studies in agriculture fields do not lead to a unanimous conclusion regarding superior performance of any of the two measurement modes (16–19). Although there is a general preference towards transmittance measurements when small concentrations need to be measured, differences arise from a combination of factors such as selected wavelength range, instrument and sample characteristics, data processing/analysis, and sampling procedure (14, 15, 20, 21). Due to the reduced flexibility and versatility of measurements by transmittance in sample presentation and characteristics, in-line monitoring, remote sensing, and field applications have been leaded by NIR diffuse reflectance spectroscopy (2, 22).

NIR INSTRUMENTATION: SPECTROPHOTOMETERS

Despite proprietary instrument conformations, any commercial NIR spectrophotometer has five basic sections further detailed: (1) sample compartment, (2) light source, (3) light wave selection system, (4) detector/s, and (5) signal processor or computer. Figure 1 shows the schematic of four of the most common optical bench arrangements in conventional NIR instruments. Note how A, B, and C conformations select the light before it reaches the sample, while conformation D selects it on the reflected/transmitted light after hitting the sample.

Sample Compartment

Instruments working by reflectance do not need sample confinement for in-line measurements, but it is common to use open sample cups or sample cells confined by silica or quartz (materials transparent to NIR light) in laboratory instrumentation. Transmission instruments may work with confined sample cells as well, but with specific pre-set pathlenghts ranging from 0.1 to 10 cm, depending on the product to be analyzed (23). An integrated adjustable sample compartment with automatic flushing is used for whole grain analyzers. One of the advantages of NIR light is its ability to pass through optical glass fibers preserving most of the signal integrity (losses lower than 5% per km of cable), even if the resulting output intensity is low. This is especially useful for measurements to be made far from the physical instrument and for multiple sampling/sequential analyses in multi-plexer systems. The use of optic fibers with probes for either transmission or diffuse reflectance measurements allows sampling by immersion in liquids for controlling fermentation or other liquid reaction processes (24–26), contact on small sample areas such as works of art (27), in-vivo medical analysis (28), and development of smaller spectrophotometers (29).

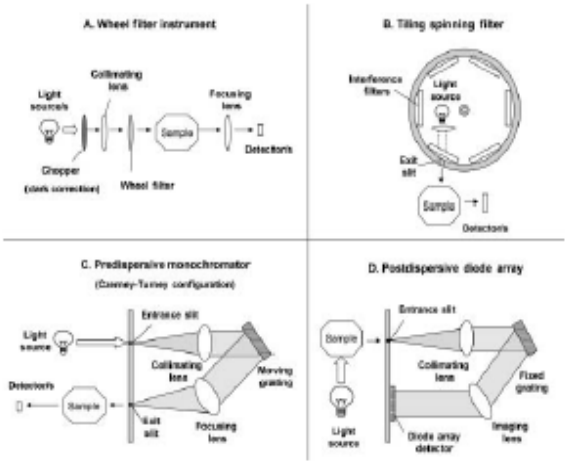


FIG. 1. Four traditional instrument conformations in NIRS. (A) Filter instrument with wheel filter. (B) Tiling spinning interference filter instrument. (C) Pre-dispersive monochromator with grating (Czerney-Turney configuration). (D) Post-dispersive diode array.

Light Sources

The most popular NIR light source is the tungsten halogen lamp, which has wavelength emission ranges from 320 to 2500 nm. The halogen gas allows recycling of the evaporated tungsten (30), and brings the advantage of a longer lifetime compared to traditional tungsten lamps without halogen.

Light emitting diodes (LED) were used as light source in the first commercial instrument for whole seed analysis in 1985 and in the first portable spectrometers (31). The low power consumption, price, small size, and long lifetime (around 25 years) of LEDs still make them the most suitable light sources for miniaturized instruments and specific screening applications outside the laboratory environment (30, 32). Conventional LEDs emit in short wavelength ranges (30–50 nm) around their center point. Several of them can be mounted in an array with narrowband interference filters if wider wavelength ranges need to be covered, although measuring many wavelengths with this configuration is not an economical approach (33). LED devices have been improved during recent years to overcome some of their limitations. For instance, some commercial instruments allow easy switching of LEDs according to the application.

Finally, the most innovative light sources are tunable diode lasers, also called superluminiscent light-emiting diodes

(SLED). Using the semi-conductor technology of diodes, tunable diode lasers are much smaller than the traditional tunable laser, cheaper, have excellent wavelength resolution, brighter, and have lower noise frequencies than tungsten lamps. SLEDs are suitable for measuring weak absorptions at good signal-to-noise ratio and as light sources in miniature instruments (34). Improvement of tunable diode lasers allows, controlling emitted light at a specific wavelength, combining light source, and wavelength selection features.

Wavelength Selection

Most detectors collect light intensity from a relatively wide range of wavelengths. Recording signal values at specific wavelengths is required for analytical purposes. Discrete wavelength values are obtained by filtering the polychromic light beam. Most simple filters work by absorption (absorption filters), which are discrete bandpass filters that absorb all light wavelengths but the one of interest. Narrow bandpass interference filters (Fabry-Perot) achieve better spectral resolution and higher output intensity by selecting wavelengths according the refractive index and thickness of the dielectric material between the two layers of reflective material (9). To select multiple wavelengths, interference filters are mounted in a wheel which can be automatically controlled to rotate and select the suitable filter for the wavelength selected (Fig. 1A and B). This creates spectrometers that provide few spectral measurements. Although filters are an alternative that provides acceptable results, problems of image misalignment and slow operation are common. (35).

Acousto-optic tunable filters (AOTF) and liquid crystal tunable filters (LCTF) allow faster tuning for wavelength selection, and provide better reproducibility without the need for mechanical devices because one filter can create several wavelengths. AOTF filters modulate the light wavelength and intensity through the interaction of sound waves generated in a birefringent TeO_2 crystal. The frequency of the acoustic signal makes the refractive properties of the crystal change allowing wavelength specific transmission. Wavelength discrimination in liquid crystal tunable filters (LCTF) is carried out by applying variable voltage to progressively change the polarity of a liquid crystal (36). Those filters provide a better output quality compared to AOTF filters, but their short wavelength range is limited (below 1800 nm), and give a lower intensity dependent on the selected wavelength (30, 35).

Dispersive-type instruments use a prism or a grating, which diffracts the incident collimated light beam at different degrees while resolving it in discrete wavelengths. Light dispersion can be done before scanning a sample (pre-dispersive instruments) or after radiating the sample with polychromatic light (post-dispersive). Post-dispersive instruments offer advantages such as less environmental interferences with the lamp radiation, analyzing wider sample areas, and holding longer distances between sample and light sources (37, 38). Prisms have been replaced by gratings because of lower cost and better linear wavelength dispersion of the last ones. There are two types: holographic (photosensitive film with fringes) and ruled (concave surface with fringes). Ruled gratings require being complemented with other optical elements such as lens, and show less stray-light rejection than holographic gratings (39, 40).

In the dispersive instruments group, there are monochromators and spectrographs such as diode-array instruments. Monochromators are pre-dispersive instruments that scan a sample with grating mechanical motion. The basic principle is as follows (Fig. 1C): polychromatic NIR light enters through an entrance slit and is then collimated (light rays are made parallel) by a mirror. The light hits the dispersion grating and later hits a focusing mirror, which reflects it to a second exit slit to either hit the sample (transmittance mode) or hit the single-channel detector (reflectance mode). Entrance and exit slits of a monochromator are very carefully designed to have accurate geometry since they are critical for instrument-observed resolution (smallest wavelength difference distinguished by the spectrometer) and effective wavelength bandwidth (full width of a band at half of its maximum value, FWHM). When using grating alone without slits, the resulting resolution is not enough for most chemical measurements in plastic or pharmaceutical applications (41). Small slits (around 0.1 mm) give low band width, more dispersion, and high spectral definition useful in qualitative applications; large slits (around 2 mm) give more intense radiation and are more suitable for quantitative analysis (39).

Diode array spectrographs are post-dispersive instruments that measure all the wavelengths at the same time thanks to a fixed grating and a set of detectors placed in array (multi-channel detectors) (Fig. 1D). There is no need for exit slits. There are fewer optical elements compared to monochromators and resolution depends on the number of elements in the detector array and array characteristics. The latest advances in wavelength selection besides tunable light sources are the micro-electromechanical systems (MEMS) created with semi-conductor technologies. MEMS diffraction gratings control light diffraction by electronically controlled movement of diffracting microelements. Their small size and lower cost has lead to a new generation of portable instruments.

Detectors

Detectors transform the incident light energy to electric analog signal. The electrical signal is then amplified and transformed to digital, which may later be further processed by the computer. Detectors and amplifiers are considered the most common sources of non-systematic noise in instruments (random noise). Random noise is reduced in most commercial instrumentation by averaging several spectra from a same sample, improving the signal-to-noise ratio (SNR). SNR achievable values in NIR spectroscopy according to Workman and Weyer range from 25,000:1 to 100,000:1 (3).

An effective detector must have a linear relationship between the energy input and signal output within its dynamic or working range—from the minimum detectable signal to the maximum before reaching saturation Measurement linearity is influenced

by other factors besides detector characteristics; for instance, the number of bits of the analog to digital converter device and slight detector misalignments, which can lead to capturing a small fraction of the reflected specular component (often called stray light) in reflectance mode instruments. Without linearity, more complex and potentially unstable mathematics are needed to calibrate the instrument.

Photo-sensitive detector materials are chosen according to the NIR region to be covered. From 400 to 1100 nm, silicon detectors (Si) are common (30). Si detectors are stable, fast, not too expensive, and sensitive to low light intensity to achieve good performance. Lead sulfide (PbS) or indium gallium arsenide (InGaAs) detectors can cover higher wavelength regions than Si detectors, being usual having both types combined in a same instrument. Photodiode array (PDAs) spectrographs have a set of InGaAs detectors or charged coupled devices (CCDs) in array. While InGaAs PDAs offer high signal precision, high SNR, and less sensitivity to high light intensities when compared to CCD, CCDs have higher signal sensitivity and resolution (42). PDAs take faster measurements (all wavelengths measured at the same time) and can be smaller in size than grating monochromators, in which the optical conformation cannot be easily reduced in size because it would lead to low throughputs and resolution (29).

Selecting Instrumentation: General Aspects

There is currently a wide range of instruments with a wide range of prices in the market: small portable instruments for little over $8,000 and big sophisticated laboratory instruments over $50,000. Instrument price increases with instrument complexity and market position. This need not mean that the most expensive spectrophotometers will lead to better performances; indeed, the opposite may be true if no further considerations are taken before purchasing analytical instrumentation. The important point is to know what the instrument function will be and what it could become in future projections. It should be taken into account that calibration cost increases with instrument cost, and the success of any NIR analytical application is highly dependent on data analysis and calibration development up to the point that instrumentation may become a relative afterthought.

To select a suitable instrument, the user must describe the nature of the materials to be tested (sample physical and chemical properties), identify potential applications or uses (environmental conditions and variability in sampling procedures), and determine the accuracy required for the analysis (i.e., screening or demanding quality purposes). Those points should be written down before looking at instruments. Instrument versatility is a relevant aspect for researchers and for users whose samples show variable composition and physical characteristics. Sampling speed, although usually not a major limiting factor in NIR spectroscopy, must be considered for in-line analysis and process monitoring. For these last applications, grating monochromators would not be recommended as they take longer scanning times and need regular wavelength standardization due to higher number of mechanical moving parts; PDAs would be more suitable. Instrument robustness is inversely proportional to the number of moving parts and dictates its suitability for rougher environments. An example are miniaturized portable instruments which show high versatility and success in applications such as material sorting for recycling purposes, screening for fraud, narcotic identification, raw material inspection, or paint thickness analysis (43).

Spectral resolution provided by manufacturers, known as observed resolution, affects spectral peak location and, hence, may impact measurement accuracies. This term can be often confused with wavelength sampling increment also in nanometers (wavelength increment between two consecutive measurements), which is greater than resolution. Although high resolution (0.1 nm) may look desirable, it is not always required for success. In analyzing biological or materials with complex composition, resolution shows low impact since the NIR absorption happens over broad regions (44, 45). Economical instrumentation with resolution over 4 nm is common and provides acceptable performances in many applications. Resolutions between 1 and 2 nm were required to obtain satisfactory discrimination of compounds with good accuracy when analyzing complex chemical matrices of pharmaceutical and mineral compounds (45, 46). Although both resolution and SNR affect instrument sensitivity and selectivity, enhancing SNR compensates for limitations caused by lower resolutions (42, 47, 48).

Technical support, periodic maintenance, and training by the supplier are important. Instrument maintenance is expensive because most operations beyond replacing the light source need to be performed by supplier personnel. Customer services availability, quality, and training are valuable support, especially during the initial stages of instrument set-up, data collection, and calibration development. Several aspects from the data acquisition software have a direct impact on instrument user-friendliness and efficiency on data managing. Current instrumentation has a wide selection of data file and calibration formats requiring varying user knowledge in data handling. Any facility in managing data is highly desirable. From the time spent since the data is collected and the calibration is developed, 80% can be spent on arranging and organizing the data (i.e., exporting, setting the right formats) and just 20% on the real data analysis (49).

Other NIR-Related Technologies

There are other NIRS technologies and instrumentation use of NIR light under slightly different principles from traditional spectroscopy. Two of the most well-established are Fourier transform NIR (FT-NIR) and NIR chemical imaging. Other emerging technologies, specifically in medical fields, such as NIR fluorescence are not discussed in this review.

Fourier transform (FT) is widely popular in MIR spectroscopy, and it has recently gained high popularity in the NIR range as well. FT technology offers advantages such as high SNR, high light outputs due to the absence of slits, fast measurements, instrumental simplicity, and high resolution and

accuracy (41). Brimmer et al. (50) claim that those advantages are more perceptible when working in the MIR region due to the limitation of higher detector noise relative to the signal when working in the NIR region.

FT-NIR measurements are carried out in time domain and the direct instrument output from sample scanning is an interferogram instead of a spectrum. NIR interferometers split the NIR light beam in two; one of the beams is reflected to a fixed mirror, and the other is reflected to a mirror that moves forward and backward at carefully controlled speed—usually tuned by a HeNe laser. The reflected beams are recombined back in the beam splitter to generate the interferogram signal, which is a result of light interferences. When displacing the moving mirror, the pathlength difference in relation to the fixed mirror changes, leading to different grades of interference between the two reflected beams and which are correlated with different light frequencies. After the interferogram light reaches the sample, the transmitted or reflected signal is read by the detector in time sequence (ms); hence, measurements are fast. Although interferograms contain information from all the frequencies or wavelengths encoded, it has to be first processed with the Fourier transform. The computation takes as an input a time domain wave signal (the interferogram) from which the transform principle states the signal is made from an addition of sinus and cosinus of a set of individual wave frequencies. The processed signal or output looks like the spectra obtained by any traditional spectrometer, but with the expectation of higher throughput and frequency accuracy. One of the drawbacks is the fact that FT-NIR instruments are complex and expensive, and mainly suitable for controlled environments (such as laboratories) due to their sensitivity to external factors such as temperature and vibrations.

Near infrared chemical imaging (NIR-CI), also called NIR hyperspectral imaging, has rapidly become popular, especially in measurements by diffuse reflectance. It combines the advantages of near infrared spectroscopy with digital mapping: the chemical compounds of a sample can be both discriminated and quantified in the sample spatial frame. This is especially useful to analyze compound distribution and sample heterogeneity. Instrument parts and operating principle are very similar to traditional spectrophotometers. The sample scanning procedure can be carried out in two ways: 1) by push-broom or moving imager technique, popular for in-line measurements and sensing, or 2) by fixed staring systems.

Pushbroom instruments measure a spectrum from a whole sample by small consecutive areas or lines while the sample platform is moved and their wavelength selection is usually by dispersion. Staring systems scan on still samples, one wavelength at a time, using either AOTF or CLTF filters. The mapping capability of imaging systems is brought by digital cameras with two dimensional arrays of detectors (pixels) such as CCDs that are effective in lower light intensities. Pixel size or area analyzed per pixel range 49 to 1,600 squared microns in commercial instruments, depending on selected magnification. Higher magnification (or smaller sample area captured per pixel) will lead to more detailed spatial analysis and a lower dilution effect of the compound of interest within the sample matrix.

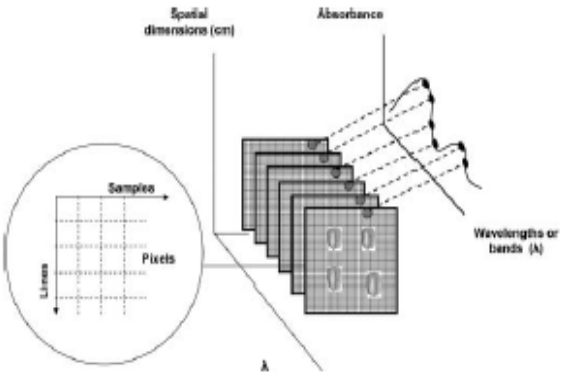


FIG. 2. NIR-CI data configuration, showing the data cube of images at different wavelengths. A single spectrum is obtained from each pixel.

NIR-CI data structure can be thought of as a cube or a stack of cards, where two spatial dimensions are combined with a third dimension corresponding to the chemical information or spectra (wavelengths). Depending on the manufacturer, around 320 × 512 pixels are arranged to capture both sample area and spectra. In that previous example, a total of 320 × 512 = 163,840 data points would be generated for a single wavelength and correlated to small sample portions as a chemical map. If the instrument had 200 sampling wavelengths, the final "image" or data cube would have a total of 320 × 512 × 200 = 32,768,000 data points. Although the amount of data generated is large, visual selection of image areas or pattern recognition techniques help in discarding pixels with no relevant information.

This concept is illustrated in fig. 2, where each squared surface is like a picture taken at one single wavelength and the small squares within represent pixels. In common imaging terminology, "samples" and "lines" specify the number of columns and rows of pixels; "bands" refer to the discrete number of wavelengths, or following the previous analogy, the number of cards in the stack.

DATA ANALYSIS AND CALIBRATION DEVELOPMENT

NIRS data analysis requires chemometrics. Chemometrics, a term widely used in NIRS-related literature, refers to the use of mathematics, statistics, and computational devices in chemical analysis. Without computing capabilities and multivariate methods, NIRS applications would not be possible. Chemometrics made possible the dealing of NIR units in resolving highly overlapped and broad peaks, high sensitivity to sample physical characteristics, and high information redundancy. Workman (3) pointed out that C-H associated vibrational information is repeated eight times in the NIR region (690 to 3000 nm). While information redundancy can be an advantage and can allow

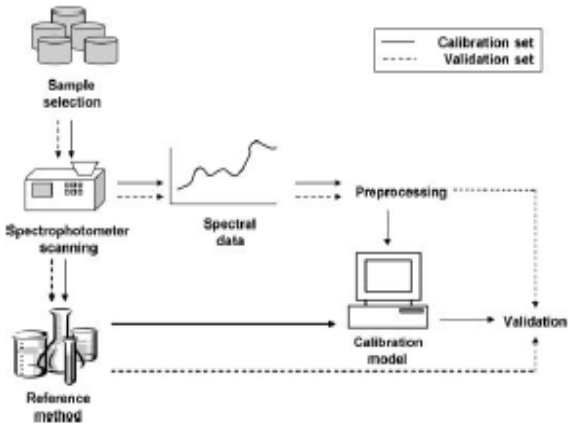


FIG. 3. Block diagram with the basic steps for NIR calibration development.

working with different wavelength ranges, determining which wavelengths hold information of interest without having correlation between them is not a problem which can be efficiently solved by trial-error experimentation.

Figure 3 shows a block diagram with the basic steps for developing a NIRS calibration. In that procedure, the broad absorptions (spectra) from a sample irradiated with NIR light are correlated with the compound concentration or sample characteristic which the user pretends to analyze. The compound to be measured should either be of organic nature (direct measurement) or be correlated with sample physical characteristics or another organic compound (indirect measurement). Some relevant aspects of the calibration procedure can be pointed out from the diagram Fig. 3: 1) there is the need for a fundamental analytical method, called the reference method, in order to obtain the dependent variable to be calibrated; 2) a suitable number of samples uniformly covering a wide enough range of analyte concentration should be part of the calibration set, and 3) the calibration model should be later validated to test the model performance on future samples.

Selecting Samples, Reference Method, and Spectral Data

The importance of choosing an adequate calibration set is often underestimated and not usually covered in the literature. There is no fixed number or rule-of-thumb to determine the number of samples to be included in a calibration. At least between 20 and 30 samples should be taken for feasibility studies and initial calibrations (51), but more robust calibrations may use a few hundred (for instance, instrument built-in calibrations for grain analysis). Calibrations of homogeneous mixtures (i.e., pharmaceutical powders) may require smaller calibration sets than agriculture samples of high compositional complexity and heterogeneity, such as whole grains or forages.

Users work under the constraints of sample availability and reduced budget. Nevertheless, there is not enough emphasis on the ultimate consequence of using calibrations developed with inadequate calibration sets: calibrations with low predictive ability. An ideal calibration set should cover the chemical, spectral, and physical characteristics of the population to be analyzed and avoid future extrapolations when predicting new samples (52). The distribution of reference values should be uniform. If the distribution is normal (bell shaped distribution), samples belonging to either higher or lower concentrations have the chance to get more relevance in the calibration, which would not be desirable.

Because reference values are not always known and reference analyses of large sample sets may be expensive, there are other methods to select an initial calibration set, using spectra. A method developed by Naes (53) and later illustrated by Naes et al. (54) uses principal component analysis (PCA) on the spectra and cluster analysis of the data.

PCA is a technique that projects the spectral data to a new reduced dimensional space. Wavelengths, the initial set of variables, are projected in the new space defined by axis called principal components (PCs). PCs are consecutively created following the directions of data variability in descending order while holding an orthogonally constraint among them. During this procedure, the initially highly redundant variables (wavelength absorbances) are substituted by a small new set of uncorrelated variables (PCs).

NIRS calibrations can match or virtually achieve better precision and accuracy than traditional wet chemistry methods (55), but paradoxically, NIRS relies on them for calibrations. The quality of the reference data influences NIRS calibrations. Careful search for the suitable method and laboratory should be carried out. In the case of NIRS instrumentation for grain analysis, calibrations are often pre-loaded, e.g., wheat protein. Although this may seem to be an opportunity to save time and resources in developing custom calibrations, the performance of any built-in calibration must be carefully validated to determine its suitability for a particular situation. Calibrations from an instrument brand and model may not perform successfully when loaded to a similar instrument, or used on different samples than the original calibration population.

Outliers from either reference values or spectral data exist and most calibration methods are highly sensitive to them (56, 57). Some tests and statistics such as Dixon test (58) or Grubbs studentized mean deviation (59) can be used as an assessment for potential outliers from the reference data a priori. For instrumental data, visual check of the spectra can identify abnormal and noisy spectra. A visual check is often not enough, and possible outliers may not be detected until data is either pre-processed, or a first attempt of calibration has been carried out.

Detecting multiple potential outliers is not simple; their effect is masked with each other. Traditional approaches to detect single outliers do not perform well (60, 61). The use of influence measures such as leverage or Hotelling's T^2 statistic in combination with checking model residuals are a powerful alternative for outlier detection (61). Influence statistics give an idea of how different a sample is from the rest of the data in a given

dimension. While that does not explicitly make a data point an outlier, high leverage followed with a high residual value (the sample was poorly modeled by the model) give chances that the sample is an influential outlier. Its exclusion from the calibration set could improve the calibration. However, if removed, enough similar samples should remain in the calibration set to avoid significant reduction of representativeness, especially in reduced data sets.

Spectra Pre-treatments

Pre-treatments or spectral pre-processing methods are a set of mathematical procedures on spectra before developing a calibration model. Mathematical pre-treatment of spectra reduces noise or background information (smoothing techniques) and increases signal from the chemical information (differentiation). Any pre-treatment must lead a robust model with good predictive ability. Basically, pre-processing methods can be classified as baseline correction—normalization, signal enhancement, and statistical filtering of signal noise. Mean centering the spectra is a basic pre-treatment that removes the absolute absorbance value (absolute baseline) and thus the need for a model intercept, and enhances the absorbance from each individual wavelength. This pre-treatment is commonly used for PCA-based calibration methods, such as partial least square regression (PLS). Centering the data to the mean value reduces the final model complexity, often reducing the number of variables to be employed by one (62).

Scaling spectra involves dividing each wavelength data by its standard deviation, which allows each wavelength to have the same weight or relevance during calibration development. Haaland and Thomas (62) suggest not using scaling when a big part of the spectra do not contain useful information because variables that have more noise than relevant information will get the same importance as the ones with relevant signal.

Multiplicative scatter correction (MSC) (63) and standard normal variate (SNV) (64) are two widely known methods that reduce spectral distortions due to scattering. SNV centers and scales each spectrum individually, so each has a mean equal to 0 and standard deviation equal to 1. MSC is more complex and memory-consuming than SNV and depends on the whole spectra set, while SNV treats each spectrum individually and independently. When applying MSC, the spectra is first averaged and each individual spectrum is regressed by partial least squares to the total average. The regression equation slope and intercept represent the additive and multiplicative effects of light scattering, respectively. Finally, each spectrum is corrected for offset (the offset value is subtracted) and each wavelength of the spectrum is divided over the slope. The regression coefficients should be stored and applied to new data. Generally, both methods provide the same results for most applications (65); this is true with all of the pre-processing methods belonging to the same category. However, MSC and SNV lead to different data geometry in the working space, where SNV shows more curvature in PCA score plots and MSC shows higher tendency to accentuate outliers (66). Pre-treatments can be very helpful but there is always a tradeoff between information loss and noise reduction: when removing scattering effects, the chemical signal may also be reduced.

The use of derivatives is an alternative for correcting the effect of overlapping peaks (enhancing signal) and removing spectral base line offset (constant drift of the spectra base line intensity across wavelengths) and baseline slope (additive variation of the spectra base line intensity across the wavelengths). The calibrations resulting from applying derivatives usually require fewer variables and models are considered to be more robust (67). Savitzky-Golay derivatives (68) are the most popular. Previously, to carry out derivation to a spectrum, a polynomial function of a selected degree is fit to a window of spectral points by least squares. This step helps smoothing the spectra or carrying out data filtering. Selecting higher polynomial degrees and small window size leads to a high function fit to the data, but the noise is modeled as well, with no smoothing effect. Low polynomial order and wide window size may lead to excessive smoothing and deletion of spectra features containing information. Subsequent derivation of the fitted polynomial differentiates overlapping signal peaks. Although it is possible to work with high-degree derivatives, most of the works in the literature use a maximum of fourth degree for curve sharpening and absorber separation. First and second derivatives are the most common and provide satisfactory results (67). First derivatives removes baseline offset while second derivatives correct the signal terms that vary linearly across the wavelengths (baseline slope) (9).

For most of the cases, the performance obtained by simple processing methods may be better than other more sophisticated methods such as orthogonal signal correction (OSC), developed by Wold et al. (69). OSC creates a model to remove any signal orthogonal (perpendicular in vector terminology) to the information of interest; thus, this method requires having previous reference data and it achieves best performances when most of the irrelevant information is actually orthogonal.

The optimum pre-treatment for a given spectra depends on the type of signal (i.e., transmittance, reflectance), sample characteristics, instrument conformation, and application or final goal (calibration or discrimination). There is no absolute or general rule for choosing the adequate pre-processing method; it usually requires a trial-error process guided by experience. Reflectance measurements often benefit from methods that reduce light scattering effects, such as MSC or SNV. Sometimes, the predictive ability of a calibration model is not improved with further mathematical treatments. Predictive ability may worsen if pre-processing excessively smoothes the signal, affecting the model ability for predicting new samples (generalization capability). Figure 4 shows the effects of several pre-processing methods on absorbance spectra of 100 bulk corn samples obtained from a reflectance mode instrument (Fig. 4, Spectra A). The offset baseline correction can be observed when preprocessing with SNV (Fig. 3, Spectra B), MSC (Fig. 3, Spectra C),

and first and second Savitzky-Golay derivatives (Fig. 3, Spectra D and E, respectively): the 100 spectra are more grouped together after eliminating scatter effects. Second Savitzky-Golay derivatives (Fig. 3, Spectra E) shows flat spectra baseline, similar to linear baseline correction (F), from which the baseline slope has been corrected. Between first and second derivatives, second derivatives lead to more peaks and an enhancement of noise on both spectral extremes. Although pre-processed spectra may look the same in the plots, differences among absorbance intensities across wavelengths still exist and are correlated with the compound to be analyzed during calibration development.

Calibration Models

NIR calibration models correlate either raw or pre-processed spectra with one or more chemical-physical property of a set of samples. As complicated as it may sound, there are several well-developed calibration methods proven to work with most of NIR applications. Those are included in all chemometric software packages. The first assumption when carrying out a calibration is the linear correlation between analyte or property to be measured and its absorbance according to Beer's law. Multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) are three of the best known calibration methods that work under this assumption. Among them, PLS is preferred for several reasons explained later.

MLR was first introduced by Strenberg et al. (70) and was the method that Norris used for his pioneer work in NIR. The method is an extension of bivariate regression for more than two variables. One of the limitations of using MLR with NIR data is the fact that does not account for wavelength multi-collinearity or variable codependency. When variables are correlated—as it is often the case with NIR wavelength—the resulting calibration is unstable and does not hold a unique solution. For this reason, the method is only appropriate if few weakly or non-correlated wavelengths are selected. Both PCR and PLS successfully deal with wavelength correlation. PCR is a direct application of the principal component analysis (PCA) method, and once the spectral data is projected to the new orthogonal non-correlated dimensional axis (PCs) a regression process by least squares is performed between the projected data and the reference values. Wold's introduction of PLS (1975) (71) was an improved alternative to PCR; both methods carry out regression on data projected to a new dimensional space, but the new space coordinates created in a process similar to PCA in PLS regression take into account the information from the reference value matrix, and PLS is thus classified as a supervised regression method; the new variables receive the name of latent variables (LVs) instead of principal components (PCs) as the new variables are not exactly the same as PCs. PCR and PLS calibrations are only based on a relatively small number of PCs/LVs because since they are extracted following the direction of maximum data variability, the last PCs/LVs usually involve noise. If an excessive number of variables are included in the calibration, a fraction of noise is also modeled and the calibration becomes too specific to the calibration set. This phenomenon is known as overfitting, and as with excessive data smoothing, it leads to a reduction of model accuracy in future predictions.

There are different approaches to estimate the appropriate number of PCs/LVs to be kept for the calibration. One of the most employed uses cross-validation, later called as an approximate validation method, and selecting the number of variables that leads to the lowest cross-validation predicted residual error sum of squares (PRESS). The PRESS cumulative function can be plotted versus the number of variables and users may visually select the first minimum from the plot. PLS accuracies may not usually be significantly higher than those of PCR but they are achieved by including fewer latent variables in the final calibration (72–74). Although disadvantages such as overfitting when reference data are noisy and higher model complexity are reported, PLS is preferred because the algorithm is faster, models have higher precision, and it provides more harmonious calibration models (75). There are several PLS-based methods in the literature (modified PLS, hybrid PLS, robust PLS) which may help to improve PLS accuracy in data sets with specific characteristics (i.e., noisy data), although they often require advanced user programming skills, that are not included in commercial chemometric software packages and the resulting models are not compatible with the instrument model files.

There may be cases where the relationship between sample spectra and reference values is not linear. Any of the previously cited calibration methods can handle small non-linearities, but when prediction residuals show certain patterns of positive and negative values or plots of predicted versus reference values show appreciable curvature, non-linearity needs to be addressed (76). Often curvature may not be noticeable but for the fact that calibration statistics are not good. The Durvin-Watson statistic can be used as an assessment tool for detecting non-linearity (77). Once the problem is detected, there are some solutions suggested by Naes et al. (76) such as new preprocessing, deleting wavelengths, adding extra principal components/latent variables to the model, using non-linear calibration models, or splitting the data in subsets using an approach similar to the cluster analysis for sample selection (78).

PCR and PLS when used with locally weighted regression can handle some cases of non-linearity. Locally weighted regression (LWR) is based on the selection of a specified number of data neighbors and applying either PCR or PLS to the local group. When predicting a new sample, all data is retrieved to find the closest data points in the working dimension, a calibration is developed with the selected local points, and the new sample is predicted with the resulting calibration model. The method is very good dealing with highly clustered data but since calibrations are done with small sets of neighbors, extreme care has to be taken to obtain stable local calibrations (79). For this reason the approach requires a higher number of samples than traditional PCR or PLS calibrations. It is recommendable to keep adding samples to the calibration pool (80).

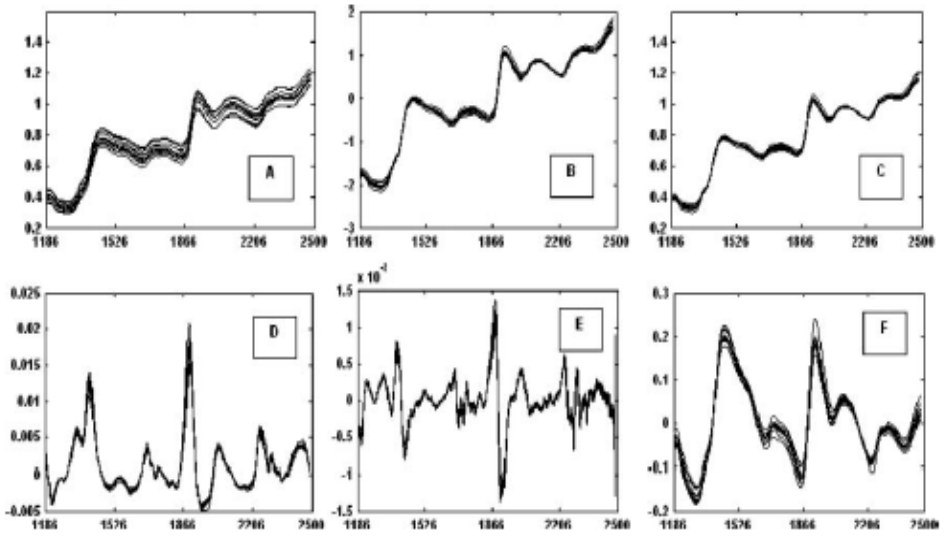


FIG. 4. (A) Raw absorbance spectra. (B) SNV Preprocessed spectra. (C) MSC preprocessed spectra. (D) First derivative spectra, window size 5 and third order polynomial. (E) Second derivative spectra, window size 5 and third order polynomial. (F) Linear baseline correction (slope correction).

Artificial neural networks (ANN) is a computational method that can be applied to NIR data to develop non-linear calibrations. By trying to simulate the human nervous system, ANN uses the calibration set to learn about any relationship that may exist between spectra and references. An artificial neural net is composed by neurons (the basic units) or nodes, layers, and transfer functions. When working with NIR spectra, the input nodes would be either wavelengths or principal components, and the output node would be the predicted value. Other nodes may be created in hidden layers (multi-layer perceptron model), which increase model complexity and the ability to model non-linear relationships. The nodes are linked by transfer functions, which are continuous functions. When the net morphology is defined (i.e., number of input nodes and hidden layer nodes), it is trained usually by backpropagation to start the learning process. In training by backpropagation algorithm, random weights are assigned to each transfer function and are updated according to the prediction error, which is propagated back through the net elements (nodes and transfer functions). This process is done numerous times (epochs or iterations). The learning rate (measurement of the change rate of weights in each epoch) and the number of epochs have to be closely monitored to check for model instability and overfitting (81). Using an additional sample set (early stopping set) besides calibration and validation sets is a common practice to avoid overfitting. Once the training is finalized, the resulting model is a function that depends on several weights coming from transfer functions, which on their turn may depend on other transfer functions and their respective weights. Therefore, the interpretation of weight meanings and the training process is rather complex. Complexity increases as more nodes and transfer functions are added. For this reason the use of ANN is not as user-friendly as other calibration methods. First, there is a high number of parameters that need to be adjusted in the net morphology and training. Second, large number of samples is needed. Third, there is a high risk of obtaining local minima solutions due to the nature of the error function involved in the training process (79).

A relatively new and more robust alternative to ANN, support vector machines (SVM), has been recently introduced for non-linear NIR calibrations, although the original concept of the method was introduced in the 1960s by Vapnik and Lerner (82) was used for linear threshold classifiers. In a later adaptation of the algorithm (83), SVM creates a tube-shaped regression volume with variable diameter. The "kernel trick" originally introduced by Aizerman et al. (84) made the algorithm very popular because it opened the opportunity of applying the linear regression algorithm in higher dimensional data because dimensionality does not matter in the final optimum SVM regression function. This basically says that while a linear correlation may not be possible in the initial dimension, the correlation may be linear in another highly dimensional combination of features. The initial data can be mapped to the higher dimensional space applying a mapping function called kernel or kernel function. There are several kernels with variable complexity [polynomial, Gaussian or radial basis function (RBF) . . .] with which users can experiment, although more complex kernels may be prone to overfitting issues (85). Few parameters such as function regularization and kernel parameters (for instance, width in the case of RBF kernel) need to be chosen for optimum prediction ability. However, the number of parameters to be adjusted is

TABLE 1
Table of Common Validation Statistics for NIR Calibrations

Statistic	Units	Equation
Coefficient of determination (r^2)	Unitless	$r^2 = \frac{\left(\sum_{i=1}^{n} \hat{y}_i y - \sum_{i=1}^{n} \hat{y}_i \sum_{i=1}^{n} y_i / n\right)^2}{\left(\sum_{i=1}^{n} \hat{y}_i^2 - \left(\sum_{i=1}^{n} \hat{y}_i\right)^2 / n\right)\left(\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n\right)}$
Standard error of prediction (SEP)	Same as reference values	$SEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i - bias)^2}{n-1}}$
Root mean square of the error of prediction (RMSEP)	Same as reference values	$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$
Bias (d)	Same as reference values	$d = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)}{n}$
Ratio of performance of deviation (RPD)	Unitless	$RPD = \frac{Sd_y}{SEP}$

$\hat{y}_i = i^{th}$ validation sample predicted value
$y_i = i^{th}$ validation sample reference value
n = number of samples in validation set
Sd_y = standard deviation of reference values from the validation set

much smaller than ANN. Other advantages of SVM over ANN are fewer samples are required and a resistance to local minima since SVM uses a Lagrangian function that has a single general minimum (86). The use of kernels is not reduced to SVM regression; it has been also proposed for ANN transfer functions and even applied in PLS or PCR. Bennet and Embrechts (87) report similar results to SVM with the advantage of simpler optimization of the regression parameters and higher model stability.

Model Validation

An adequate validation of the calibration models is a crucial step to determine the suitability of the model to predict new samples, which is the whole purpose of developing NIR calibrations. Ideally, the best validation should be done with distributed samples which were not previously used for calibrating. Since independent validation may not always be possible, cross-validation can provide a basic assessment regarding calibration performance. The general idea of the method is to keep a single sample (full cross-validation) or a group of samples (k-fold cross-validation) apart and develop a calibration with the remaining samples. The developed calibration is validated with the excluded samples and the prediction values are recorded. This procedure is consecutively done until all the samples have been predicted once. The final calibration model is not tested, but rather several submodels developed with calibration data subsets. Any statistic reported from cross-validation can not be directly compared or interpreted the same way as statistics from a real validation of the final model with new samples. The standard errors from cross-validation are often optimistic and, especially in k-fold validation, highly affected by data artifacts (88). However, reporting cross-validation statistics are preferred over reporting calibration results alone.

Table 1 shows the most used NIR validation statistics among those suggested and detailed in (51). However, it is not unusual to find literature using other statistics, reporting not so relevant figures of merit, or simply not reporting enough information for a good statistical assessment of the model quality. The coefficient of determination (R^2), which provides an estimation of how much variance between reference and predicted values is explained versus the total variance, seems to be one of the erroneously preferred guides for validation assessment. Its high dependency on the reference value range is often ignored (89). The standard error of prediction (SEP, or SECV when reporting cross-validation results) provides information regarding calibration precision. SEP is corrected for the bias value (or systematic error); thus, when reporting SEP bias must be reported as well. The square root of mean standard error of prediction (RMSEP) is related to SEP and bias according to (Eq. 3). Because RMSEP accounts for bias and provides information regarding calibration accuracy, it can be reported alone, especially when bias is small (then RMSEP ~ SEP) (90).

$$RMSEP^2 = SEP^2 + Bias^2 \quad [3]$$

The final statistic to be discussed is the ratio of performance of deviation or relative predictive determinant (RPD), which is dimensionless and specific to NIR spectroscopy. It is related with the ability of the model to predict future data in relation to the initial variability of the calibration data. Basically, if a calibration leads to a low SEP but the calibration was carried out with a small range of reference values (standard deviation of reference values almost the same as SEP), the model would only

be predicting the data average. Williams (51) provides ranges of RPD values related to the calibration suitability: values above 8 indicate that the calibration can be used for any purpose, while values below 2.3 indicate a poor calibration performance, with use for predicting new samples not advisable.

FINAL REMARKS

Near infrared technologies offer fast solutions for organic compound discrimination and quantification. With the instrumental market in constant growth and development, cheaper and yet more accurate instruments will probably offer opportunities to explore new applications and fields of work. But choosing a suitable instrument for an application involving the use of NIRS is not even half of the requirements for its success. Sample selection, chemometric methods, and validation are key factors that should not be overlooked. Far from discouraging new NIRS users, this review intended to point out those critical steps and warn for extra care. Although the main focus was on NIRS quantitative analysis, the mentioned steps and methods in this review can be easily adapted for discriminative analysis, and the critical stages remain the same. Although it may have not been especially emphasized in the review, we wish to offer advice about keeping any calibration in constant update, especially when working with samples that may suffer any kind of periodic or seasonal change, and be aware of any change in the reference method as it will negatively impact the calibration. Keeping the sample calibration pool updated by adding new samples when required and using standards for periodic checks is a must to keep a good predictive ability in NIRS calibrations over time. Standardization tasks may be an other periodically unavoidable task since instrument drift and/or environmental changes are not completely under control and impact instrument performance. Readers can refer to several literature sources to learn about keeping the calibration and instrument performance up to date (91, 92).

LIST OF ABBREVIATIONS USED IN THE PAPER

AACC	American Association of Cereal Chemists
ANN	Artificial neural networks
AOTF	Acousto-optic tunable filters
CCD	Charged couple device
FIR	Far infrared
FT	Fourier transform
FWHM	Full width at half maximum
LCTF	Liquid crystal tunable filters
LED	Light emitting diodes
LV	Latent variables
LWR	Locally weighted regression
MEMS	Micro-electro-mechanical systems
MIR	Middle infrared
MLR	Multiple linear regression
MSC	Multiplicative scatter correction
NIR	Near infrared
NIR-CI	Near infrared chemical imaging
NIRS	Near infrared spectroscopy
OSC	Orthogonal signal correction
PCA	Principal component analysis
PCR	Principal component regression
PDA	Photo diode array
PLS	Partial least squares
PRESS	Predicted residual error sum of squares
RBF	Radial basis function
RMSEP	Root mean square of the standard error of prediction
RPD	Relative predictive determinant
SECV	Standard error of cross validation
SEP	Standard error of prediction
SLED	Superluminiscent light emitting diodes
SNR	Signal to noise ratio
SNV	Standard normal variate
SVM	Support vector machines

REFERENCES

1. F. W. Herschel, Investigation of the powers of the prismatic colours to heat and illuminate objects. *Philosophical Transactions of the Royal Society of London* 90 (1800):255–329.
2. G. Mc. Dryden, "Near Infrared reflectance spectroscopy: Applications in deer nutrition" *Publication No W03/007, Project No UQ 109A* (2003), <http://www.rirdc.gov.au/reports/DEE/w03-007.pdf>, accessed February 2009.
3. J. J. Workman, "An introduction to near infrared spectroscopy" (2005), (http://www.spectroscopynow.com/coi/cda/detail.cda?id=1881&type=EducationFeature&chId=2&page=1) accessed January 2010.
4. L. Bokobza, Near Infrared spectroscopy. *Journal of Near Infrared Spectroscopy* 6 (1998):3–17.
5. J. R. Hart, and K. H. Norris, Determination of the moisture content of seeds by near-infrared spectophotometry of their methanol extracts. *Cereal Chemistry* 39 (1962):94–99.
6. K. H. Norris and J. R. Hart, Direct spectrophotospectrometric determination of moisture content of grain and seeds. *International Symposium Washington 1963,* 4 (1965):19–25.
7. AACC International, *Approved Methods of the American Association of Cereal Chemists* (AACC, St. Paul, MN, 1999).
8. J. Workman and L. Weyer, History of near infrared (NIR) applications, in *Practical Guide to Interpretive Near-Infrared Spectroscopy* eds. J. Workman and L. Weyer (CRC Press, Boca Raton, FL, 2007), Ch. 13, 107–111.
9. N. Pou Saboya, *Analisis de control de preparados farmaceuticos mediante espectroscopia en el infrarojo proximo*. Ph.D dissertation Universitat de Barcelona (UAB), Barcelona, Spain, 2002.
10. D. J. Dham and K. D. Dham, The physics of near-infrared scattering, in *Near-Infrared Technology in the Agricultural and Food Industries* eds. P. C. Williams and K. Norris (AACC, St Paul, MN, 2001), Ch, 1, 1–17.
11. O. Berntsson, L-G. Danielsson, and S. Folestad, Estimation of effective sample size when analysing powders with diffuse

reflectance near-infrared spectrometry. *Analytica Chemica Acta* 364(1–3) (1998):243–251.

12. K. Norris and P. C. Williams, 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 Chemistry* 61 (1984):158–165.
13. P. Corti, G. Ceramelli, E. Dreassi, and S. Mattiim, Near infrared transmittance analysis for the assay of solid pharmaceutical dosage forms. *The Analyst* 124 (1999):755–758.
14. S. E. Kays, N. Shimizu, F. E. Barton II, and K. Ohtsubo, Near-infrared transmission and reflectance spectroscopy for the determination of dietary fiber in barley cultivars. *Crop Science* 45 (2005):2307–2311.
15. S. M. Short, R. P. Cogdill, and C. A. Anderson, Figures of merit comparison of reflectance and transmittance near-infrared methods for the prediction of constituent concentrations in pharmaceutical compacts. *Journal of Pharmaceutical Innovation* 3(1) (2008): 41–50.
16. B. A. Orman and R. A. Schuman, Jr., Comparison of near-infrared spectroscopy calibration methods for the prediction of protein, oil, and starch in maize grain. *Journal of Agricultural and Food Chemistry* 39(5) (1991):883–886.
17. P. C. Williams and D. C. Sovering, Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. *Journal of Near Infrared Spectroscopy* 1 (1993):25–32.
18. T. Börjesson, B. Stenberg, and J. Schnürer, Near-infrared spectroscopy for estimation of ergosterol content in barley: A comparison between reflectance and transmittance techniques. *Cereal Chemistry* 84(3) (2007):231–236.
19. J. Xing and D. Guyer, Comparison of transmittance and reflectance to detect insect infestation in Montmorency tart cherry. *Computers and Electronics in Agriculture* 64(2) (2008):194–201.
20. S. R. Delwiche, Single wheat kernel analysis by near-infrared transmittance: protein content. *Cereal Chemistry* 72(1) (1995): 11–16.
21. R. P. Cogdill, C. A. Anderson, M. Delgado-Lopez, D. Molseed, R. Bolton, T. Herkert, A. M. Afnan, and J. K. Drennen III, Process analytical technology case study part I: Feasibility studies for quantitative near-infrared method development. *AAPS PharmSciTech* 6(2) (2007):E262–E272
22. D. Fischer and E. Pigorsch, New developments in process control by spectroscopic methods in the polymer and plastics industry—Near infrared miniature spectrometer and high temperature and pressure near infrared and raman probes. *Paper presented at the third annual UNESCO school IUPAC conference on Macromolecules and materials science, Matieland, South Africa* (2000), (*academic.sun.ac.za/unesco/PolymerED2000/Conf2000/Fischer.pdf*). accessed January 2010.
23. E. D. Lipp, Near-infrared spectroscopy of silicon-containing materials. *Applied Spectroscopy Reviews* 27(4) (1992):385–408.
24. R. R. Buchanan, D. E. Honigs, C. J. Lee, and W. Roth, Detection of ethanol in wines using optical-fiber measurements and near-infrared analysis. *Applied Spectroscopy* 42 (1988):1106–1111.
25. E. Tamburini, G. Vaccari, S. Tosi, and A. Trilli, Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fiber probe. *Applied Spectroscopy* 57(2) (2003):132–138.
26. M. C. Sarraguca, A. Paulo, M. M. Alves, A. M. A. Dias, J. A. Lopes, and E. C. Ferreira, Quantitative monitoring of an activated sludge reactor using on-line UV-visible and near-infrared spectroscopy. *Analytical and Bioanalytical Chemistry* 395(4) (2009):1159–1166.
27. M. Bacci, C. Bellucci, C. Cucci, C. Frosinini, M. Picollo, S. Porcinai, and B. Radicati, Fiber optics reflectance spectroscopy in the entire VIS-IR Range: A powerful tool for the non-invasive characterization of paintings, in *Materials Issues in Art and Archaeology VII* eds. P. B. Vandiver, J. L. Mass, and A. Murray, (Materials Research Society, Warrendale, PA, USA, 2005), 852:297–302.
28. B. Yu, E. S. Burnside, G. A. Sisney, J. M. Harter, Z. Changfang, A. Dhalla, and N. Ramunjam, Feasibility of near-infrared diffuse optical spectroscopy on patients undergoing image guided core-needle biopsy. *Optics Express* 15(12) (2007):7335–7350.
29. J. P. Smith, Product review: Spectrometers get small. Miniature spectrometers rival benchtop instruments. *Analytical Chemistry* 72(19) (2000):653A–658A.
30. E. Stark and K. Luchter, NIR instrumentation technology. *NIR News* 16(7) (2005):13–16.
31. W. F. McClure, D. Moody, D. L. Standfield, and O. Kinoshita, Hand-held NIR spectrometry. Part II: An economical no-moving parts spectrometer for measuring chlorophyll and moisture. *Applied Spectroscopy* 56(6) (2002):720–724.
32. Axun Technologies, Designing a miniature spectrometer, *Technical note* (2005), <www.axsun.com/.../05-02-084a%20_designing_miniature_spectrometer%20Final.pdf>, accessed January 2010.
33. J. Malinen, M. Känsäkoski, R. Rikola, and C. G. Eddison, LED-based NIR spectrometer module for hand-held and process analyzer applications. *Sensors and Actuators B: Chemical* 51(1–3) (1998):220–226.
34. M. Lang, Diodes storm the tunable laser ranks, *Photonics Tech Briefs* (1999), (http://www.ptbmagazine.com/articles/diodes0199/), accessed January 2010.
35. C. Balas, Review of biomedical optical imaging—A powerful, non-invasive, non-ionizing technology for improving in-vivo diagnosis. *Measurement Science Technology* 20 (2009):1–12.
36. Y. Garini, I. T. Young, and G. McNamara, Spectral imaging: Principles and applications. *Cytometry* 69A(8) (2006): 735–747.
37. A. W. Schumann and J. H. Meyer, Progress with the implementation of diode array near-infrared spectrometer for direct at-line analysis of sugarcane samples. *Proceedings of South Africa Sugar Technology Association* 74 (2000):122– 123.
38. W. Wang and J. Paliwal, Design and evaluation of a visible-to-near-infrared electronic slitless spectrograph. *Science Technology* 17 (2006):2698–2704.
39. S. Holler, Y. Pan, R. K. Chang, J. R. Bottiger, S. C. Hill, and D. B. Hillis, Two-dimensional angular optical scattering for the characterization of airborne microparticles. *Optics Letters* 23 (1998):1489–1491.
40. J. Domanchin and J. R. Gilchrist, Size and spectrum. *Photonics* July 2001, (2001):12–118.
41. Thermo Fisher Scientific, Advantages of fourier-transform near-infrared spectroscopy. *Application note 50771* (2006), <http://www.thermo.com/eThermo/CMA/PDFs/Articles/articlesFile_1195.pdf>, accessed January 2010.
42. C. V. Greensill and K. B. Walsh, Optimization of instrument precision and wavelength resolution for the performance of NIR

calibrations of sucrose in water–cellulose matrix. *Applied Spectroscopy* 54(3) (2000):426–430.

43. Y. Geller, A new approach to NIR spectroscopy allowing remote analysis. *LabPlus International* 20 (2006):13–16.
44. P. R. Amstrong, F. X. Maghirang, and F. E. Dowell, Comparison of dispersive and Fourier-transform NIR instruments for measuring grain and flour attributes. *Applied Engineering in Agriculture* 22(3) (2006):453–457.
45. L. Mcarthur and C. Greensill, Impact of resolution on NIR PLS calibration of kaolinite content with weipa bauxite. *Measurement Science Technology* 18 (2007):1343–1347.
46. H. Chung, S., Choi, J., Choo, and Y. Lee, Investigation of partial least squares (PLS) calibration performance based on different resolutions of near infrared spectra. *Bulletin of Korean Chemistry Society* 25(5):647–651.
47. T. Tarumi, A. K. Amerov, M. A. Arnold, and G. W. Small, Design considerations for near-infrared filter photometry: Effects of noise sources and selectivity. *Applied Spectroscopy* 63(6) (2009): 700–708.
48. K. B. Wash, J. A. Guthire, and J. W. Burney, Application of commercially available, low-cost, miniaturized NIR spectrometers to the assessment of the sugar content of intact fruit. *Australian Journal of Plant Physiology* 27(12) (2000):1175–1186.
49. C. R. Hurburgh, Jr. and G. R. Rippke, Calibration, standardization and validation economics. *Oral presentation at the International Diffuse Reflectance Conference (IDRC)* (2008), Chambersburg, PA.
50. P. J. Brimmer, F. A. DeThomas, and J. W. Hall, Method development and implementation of near-infrared spectroscopy in industrial manufacturing processes, *in Near-Infrared Technology in the Agricultural and Food Industries* eds. P. C. Williams and K. Norris (AACC, St. Paul, MN), Ch. 11, (2001):199–214.
51. P. C. Williams, Implementation of Near Infrared technology, in *Near-Infrared Technology in the Agricultural and Food Industries*, eds. P. C. Williams and K. Norris (AACC, St Paul, MN, 2001), Ch. 8, 145–170.
52. T. Fearn, Chemometrics: An enabling tool for NIR. *NIR News* 16(7) (2005):17–19.
53. T. Naes, The design of calibration in near reflectance analysis by clustering. *Journal of Chemometrics* 1 (1987):121–134.
54. T. Naes, T. Isaksson, T. Fearn, and T. Davies, Selection of samples for calibration, in *A User-Friendly Guide to Multivariate Calibration and Classification* (NIR Publications, Chichester, UK, 2002), Ch. 15, 191–195.
55. D. B. Coats, Is near infrared spectroscopy only as good as the laboratory reference values? An empirical approach. *Spectroscopy Europe* 14(4) (2002):24–26.
56. I. V. Kovalenko, G. R. Rippke, and C. R. Hurburgh, Determination of amino acid composition of soybeans (Glycine max) by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry* 54 (2006):3485–3491.
57. M. Hubert, P. J. Rousseeuw, and S. Van Aelst, High-breakdown robust multivariate methods. *Statistical Science* 23(1) (2008): 92–119.
58. W. J. Dixon, Analysis of extreme values. *Annals of Mathematic Statistics* 21 (1950):488–506.
59. F. E. Grubbs, Sample criteria for testing outlying observations. *Annals of Mathematic Statistics* 21 (1950):27–58.
60. B. Walczak and D. L. Massart, Multiple outlier detection revisited. *Chemometrics and Intelligent Laboratory Systems* 41(1) (1998):1–15.
61. T. Naes, T. Isaksson, T. Fearn, and T. Davies, Outlier detection, in *A User-Friendly Guide to Multivariate Calibration and Classification* (NIR Publications, Chichester, UK 2002), Ch. 14, 177–189.
62. D. M. Haaland and E. V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* 60 (1988):1193–1202.
63. P. Geladi, D. MacDougall, and H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy* 39 (1985):491–500.
64. R. J. Barnes, M. S. Dhanoa, and S. J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* 43 (1989):72–777.
65. M. S. Danoa, S. J. Lister, R. Sanderson, and R. J. Barnes, The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *Journal of Near Infrared Spectroscopy* 2 (1994):43–47.
66. T. Fearn, C. Riccicoli, A. Garrido-Baro, and E. Guerrero-Ginel, On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems* 96 (2009):22–26.
67. D. W. Hopkins, Using data pretreatments effectively. *Seminar at International Diffuse Reflectance Conference* (2008), Chambersburg, PA.
68. A. Savitzky and M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8) (1964):1627–1639.
69. S. Wold, H. Antti, F. Lindgren, and J. Öhman, Orthogonal signal correction of near-infrared spectra. *Chemometric Intelligent Laboratory Systems* 44 (1998):175–185.
70. J. C. Sternberg, H. S. Stillo, and R. H. Schwendeman, Spectrophotometric analysis of multicomponent systems using the least squares method in matrix form. *Analytical Chemistry* 32 (1960):84–90.
71. S. Wold, Soft modeling by latent variables; the non-linear iterative partial least squares approach. in *Perspectives in Probability and Statistics, Papers in Honour of M.S. Barlett*, ed. J. Gani (Academic Press London 1975).
72. T. Naes, C. Irgens, and H. Martens, Comparison of linear statistical methods for calibration of NIR instruments. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 35(2) (1986):195–206.
73. B. Hammateenejad, M. Akhind, and F. Samar, A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: Effect of wavelength selection. *Spectrochimica Acta, part A: Molecular and Biomolecular Spectroscopy* 67(3–4) (2007):958–965.
74. R. Muñiz, M. A. Perez, C. De la Torre, C. E. Carleos, N. Corral, and J. A. Baro, Comparison of principal component regression (PCR) and partial least square (PLS) methods in prediction of raw milk composition by VIS-NIR spectrometry. Application to development of on-line sensors for fat, protein and lactose contents. Oral presentation proceedings for XIX IMEXO world congress of applied Metrology, Lisbon, Portugal (2009). Viewed January 2010. http://www.imeko2009.it.pt/Papers/FP_229.pdf.

75. J. H. Kalivas and P. J. Gemperline, Calibration, in *Practical Guide to Chemometrics*, ed. P. J. Gemperline (CRC, FL, Taylor and Francis Group, Boca Raton 2006) Ch. 6, 105–166.
76. T. Naes, T. Isaksson, T. Fearn, and T. Davies, Non-linearity problems in calibration, in *A User-Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, UK, 2002, Ch. 9, 93–97.
77. M. Howard and J. Workman, Jr., Chemometrics in spectroscopy-Linearity in calibration: The Durbin-Watson statistic. *Spectroscopy Magazine*, 20(3) (2005):34–40.
78. T. Naes, Multivariate calibration when data are split into subsets. *Journal of Chemometrics* 5 (1991):487–501.
79. F. Despagne and L. Massart, Neural networks in multivariate calibration. *Analyst* 123 (1998):157R–178R.
80. D. A. Burns and E. W. Ciurczak, *Handbook of Near-Infrared Analysis* (CRC Press, Taylor and Francis Group, Boca Raton, FL, 2007).
81. B. G. M. Vandegiste, D. L. Massart, L. M. C. Buydens, S. De Jong, P.J. Lewi, and J. Smeyers–Verbeke, Artificial neural Networks, in *Handbook of Chemometrics and Qualimetrics: Part B*, eds. B.G.M. Vandeginste and S.C. Rutan (Elsevier, Amsterdam, 1998), Ch. 44, 649–695.
82. V. Vapnik and A. Lerner, Pattern recognition using generalized portrait method. *Automation and Remote Control* 24 (1963):774–780.
83. H. Drucker, C. J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik, Support vector regression machines. *Advances in Neural Information Processing Systems* 9 (1997):155–161.
84. M. Aizerman, E. Braverman, and L. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* 25 (1964):821–837.
85. O. Ivanciuc, Applications of support vector machines in chemistry. *Reviews in Computational Chemistry* 23 (2007):291–400.
86. S. Zomer, Classification with support vector machines. *Homepage of Chemometrics* (2004), <http://www.chemometrics.se/images/stories/pdf/nov2004.pdf>, accessed October 2009.
87. K. P. Bennett and M. J. Embrechts, An optimization perspective on kernel partial least squares regression, in *Advances in Learning Theory: Methods, Models and Applications* NATO Science Series III:Computer & Systems Sciences, ed. J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle (IOS Press, Amsterdam, 2003), Vol. 190, 227–250.
88. T. Naes, T. Isaksson, T. Fearn, and T. Davies, Validation, in *A User-Friendly Guide to Multivariate Calibration and Classification* (NIR Publications, Chichester, UK, 2002), Ch. 13, 155–160.
89. T. Fearn, Assessing calibrations: SEP, RPD, RER, and R^2. *NIR News* 13(6) (2002):12–14.
90. A. M. C. Davies and T. Fearn, Back to basics: Calibration statistics. *Spectroscopy Europe* 18(2) (2006):31–32.
91. C. N. G. Scotter, NIR techniques in cereal analysis, in *Cereals and Cereal Products: Chemistry and Technology* eds. D. A. V. Dendy and B. J. Dobrasky (Aspen Publishers, Gaithersburg, MD, 2001), Ch. 5, 90–98.
92. E. Bouveresse and B. Campbell, Transfer of multivariate calibration models based on near-infrared spectroscopy, in *Handbook of Near Infrared Analysis* eds. D. A. Burns and E. W. Ciurczak (CRC Press, Boca Raton, FL, 2007), Ch. 11, 231– 244.

APPENDIX 2. TRAINING MANUAL OF THE GRAIN QUALITY LABORATORY

Near Infrared Spectroscopy (NIRS) and Chemometrics

Training Manual

Lidia Esteve Agelet
Benoit Igne
Igor Kovalenko

Grain Quality Laboratory
Department of Agricultural and Biosystems Engineering
Iowa State University

2005 – 2011

PREFACE

The purpose of this manual is to introduce the basics of near-infrared spectroscopy (theory, terminology, instrumentation, and applications) and calibration development to beginners. The authors intend to provide the key aspects of NIRS to ultimately guide novices through the process of understanding the critical aspects of this technology and developing robust calibrations. Through the use of user-friendly approaches with limited chemical, mathematical terminologies and equations, new users with a basic background in sciences should be able to gain a general but relevant view of the technology. The steps to develop applications can be mastered with the examples, data, and exercises provided in this manual so the reader is ready to work on their own applications.

TABLE OF CONTENTS

1. INTRODUCTION AND PRINCIPLES OF NEAR INFRARED SPECTROSCOPY 1
1.1. What is NIR spectroscopy? 1
1.2. Properties of radiation 1
1.3. Principles behind absorption of NIR 3
1.4. Using NIR spectra for analytic purposes 6
1.4.1. Principles of Beer's law 6
1.4.2. Measuring absorption of radiation..... 8
1.4.3. The NIR spectra 11

2. HISTORY OF NEAR INFRARED SPECTROSCOPY 15

3. APPLICATIONS 16

4. INSTRUMENTATION 19
4.1. Sample compartment 20
4.2. Light sources 21
4.3. wavelength selection 22
4.4 Detectors 27
4.5 Selecting instrumentation: general aspects 29
4.6 List of popular manufacturers of NIR spectrometers 31

5. OTHER NIRS-RELATED TECHNOLOGIES 32
5.1. Fourier-transform Near Infrared spectroscopy 33
5.2. Near Infrared chemical imaging 34

6. CHEMOMETRICS AND CALIBRATION PROCESS 36
6.1. Selecting samples and reference method 39
6.2. Getting the data ready 42
6.3. Data preprocessing (pretreatment) 50
6.4. Calibration methods 56
6.4.1. MLR 57
6.4.2. PLS and PCR 58
6.4.3. ANN 62
6.4.4. SVM 65
6.5. Validation 67

7. ADVANCED TOPICS 72

7.1 Optimization methods 72
7.1.1 Variable selection 73
7.1.2 Local regression 78
7.2 Standarization 83
7.2.1. Optical standardization techniques 83
7.2.2 Post regression correction techniques 90
7.2.3. Robust standardization techniques 92

8. THE UNSCRAMBLER EXAMPLES 97
8.1. Carrying out Pricnipal Component Analysis 97
8.1.1 Importing the data 97
8.1.2. Ploting the data 98
8.1.3 Carrying out PCA 99
8.1.4 Checking the results 102
8.1.5 Saving the model and applying it 106
8.2. Developing a basic PLS calibration 108
8.2.1 Checking the data 108
8.2.2 Calibration process and results 112
8.2.3 Saving the model and applying it 112

9. MATLAB EXAMPLES 116
9.1. Importing XLS files 116
9.2. Importing TXT files..... 118
9.3. Importing SPC files..... 119
9.4. Saving MATLAB data in TXT and CSV files 119
9.5. Storing spectral and reference data in MAT files 120
9.6. Opening MAT files in The Unscrambler 121
9.7. Calibration using PLS regression 125
9.8. Calibration using ANN regression 128
9.9. Calibration using LS-SVM regression 131
9.10. Creating uniformly distributed data sets 133
9.11. Removal of spectral outliers 135

10. RECOMMENDED READING 141
10.1. Books 141
10.2. Internet resources 142

11. REFERENCES 143

1. INTRODUCTION AND PRINCIPLES OF NEAR INFRARED SPECTROSCOPY

1.1) What is Near Infrared Spectroscopy?

Near Infrared Spectroscopy (NIRS) is one of the several technologies included in a bigger group of analytic technologies called **vibrational spectroscopy**. All of them have in common that they are based on the analytical use of the vibration that light causes to atoms or molecules (that is the reason why is called vibrational spectroscopy). The term "spectroscopy" is derived from **spectra**, which as you will later learn, is the result of recording that vibration in means of absorbed light.

As analytical technique, NIRS is used to analyze **organic compounds** based on their absorption of near infrared light. Organic compounds are those that have functional groups with O-H, N-H, C-H bounds. Explained in a more manner, compounds which can be found in organic matter or living beings. Some examples of organic compounds currently measured in grains are protein, starch, or fat. Water is not an organic compound but because of its O-H bound can be also measured by NIRS – in fact, it is one of the compounds more easily measured! -. This analysis can be **qualitative** (classify or identify. For instance, classify grains according to high or low concentration of fat) or **quantitative** (predict the concentration of amount of certain compound in a sample).

1.2) Properties of Radiation

Let's first talk about light, what is light? Light is energy and has wave properties, but it also has mass properties as it is made of small particles or energy packages called photons (this is known as the wave-particle duality). Light is radiation that can be seen. There are other non-visible radiations with variable energey. When describing radiation

in terms of energy or wave, it is common to name its **frequency** or its **wavelength** in spectroscopy. Figure 1 shows those two concepts. The wavelength is the distance between peaks of a wave. The frequency would be the number of oscillations in the wave per second, which is the definition of the unit **Hertz (Hz).** The higher the frequency, the more energetic the radiation is. It is the opposite with the wavelength: smaller wavelengths come from more energetic radiation.

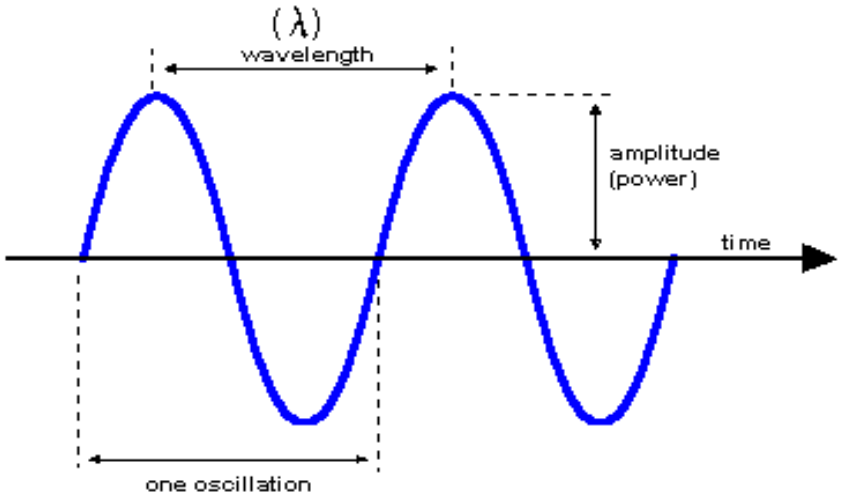


Figure 1. Principal parts of waves: wavelength, oscillation (which is related to frequency), and amplitude

When you arrange radiation in function of its energy, frequency or wavelength, you get the **electromagnetic spectra** shown in figure 2. In the figure you will recognize several names of energetic waves (micro waves and X-rays for sure!), all of them with very different properties leading to very different applications. Now give it some thought. Which radiation is more energetic, near infrared (NIR) or microwaves? If you check well the electromagnetic spectrum, the near infrared region is inside the infrared region, in grey, close to the visible light. Microwaves are next to the right of the infrared, with higher wavelength and lower frequency, so those are less energetic.

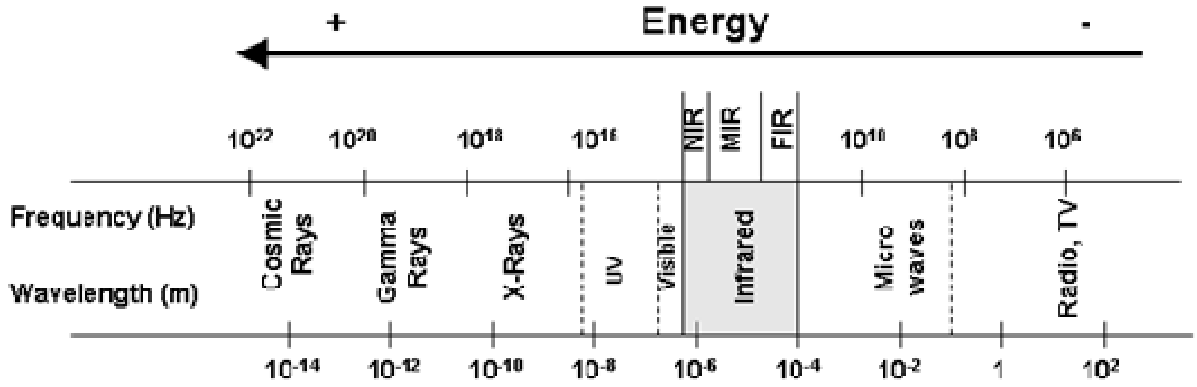


Figure 2. Electromagnetic spectra in function of frequency (Hz) and wavelength (m)

Note the wavelength units, which are international units of length: meters (m). If you remember the unit prefixes, you should be familiar with **nanometers** (1 nm = 10^{-9} m). This is the wavelength unit generally used to work in the NIR region, which covers from **750 nm to 2,500 nm**, or from 2120 to 400 THz in frequency. In the mid-infrared region and reduced communities of NIR users, a unit called **wavenumber** is also popular. A wavenumber in spectroscopy is equal to inverse centimeters (1/cm), so knowing that 1 cm is 10,000,000 nm you can convert from one unit to the other.

1.3) Principles Behind the Absorption of Near Infrared

Molecules have natural vibration at discrete frequencies according to their composition (heavy or lighter atoms) and environmental conditions such as temperature. In order for a molecule to absorb NIR radiation, it must have its molecular vibrational frequency matching the frequency of the radiation and, furthermore, the dipole momentum of the molecule must change. This last concept is more complex to explain and we will not detail it, but the concept to retain is that a specific molecule in specific environmental conditions will only be able to absorb certain frequencies or wavelengths.

Let's go a little bit deeper in the theory. The quantum theory states that molecules and atoms can only be found in states of certain discrete energy, called electronic states. There are three electronic states: ground, first, and second. In each electronic state there are other energy sublevels, called vibrational states, as you can see in figure 3. The first

level in the ground state is an equilibrium state in which the probability to find an atom or molecule is the highest in standard conditions. As represented in figure 3, we can see how more energetic radiation, such as ultraviolet, induce the molecules to jump further from the ground state, while infrared only has energy to induce to molecular vibrations. When a molecule absorbs NIR the result of this radiation-molecule interaction is stretching, bending and molecular rotations (Davies, 2005).

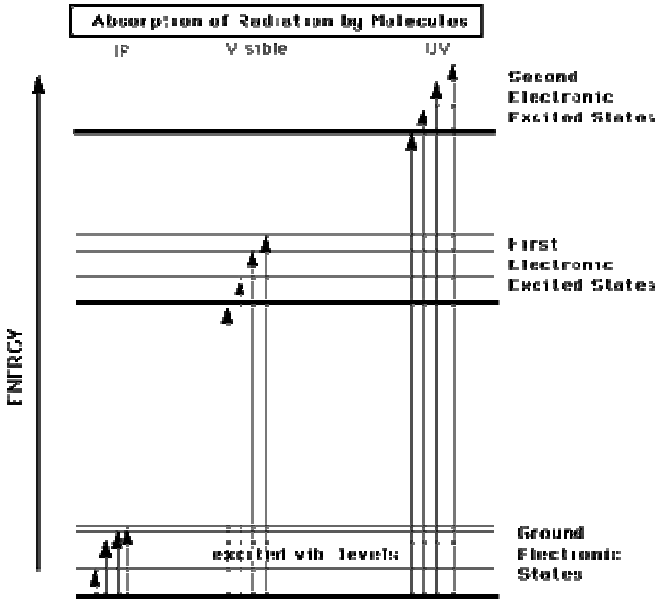


Figure 3. Diagram of molecular energy levels achieved by the absorption of light

If a molecule which is found in the ground state achieves the next one, it will need to absorb energy (from irradiated light in our case) equal to the difference between the first initial state and the new one. Because light energy as we said before depends on its wavelength and frequency, the emitted/absorbed radiation will be of a specific wavelength and frequency. In terms of molecular excitation, is there a big difference between the absorption of mid infrared and near infrared? Yes. In the beginning there were some theories that resulted being wrong. The first said that the energy between the vibrational energy levels was the same; that is to say, the states were like steps from a perfect stair, evenly spaced. This resulted not being true, you can see in figure 3 how the states are closer (less energy required) the higher the energy levels. Secondly, it was

believed that it was only possible to jump between consecutive states. When a molecule absorbs light that makes it go from one vibrational state to the next consecutive one, this is called a **fundamental absorption**. This is what happens when a molecule absorbs mid infrared light. Summarizing, the belief of the last theories made the existence of near infrared absorption not explained: if near infrared was more energetic than mid-infrared, but less energetic than UV light, then… what happened if a molecule absorbed NIR? Or, could a molecule absorb NIR if that energy did not match neither the energy for a fundamental absorption nor the energy required to go to another electronic state? It was later discovered that it is possible to jump more than one vibrational level, which phenomenon is called **overtone**. You can see this concept in figure 4. Note that each overtone requires more energy, and it is more difficult to happen. In fact, there are up to four overtones, but you will only find that literature mentions three, because the fourth one happens such as low probabilities – so few molecules absorb at those wavelengths or frequencies- that you cannot record it with the instrument, the signal is too weak. Therefore, lower wavelengths in the NIR region cause higher overtones, because they are energetic.

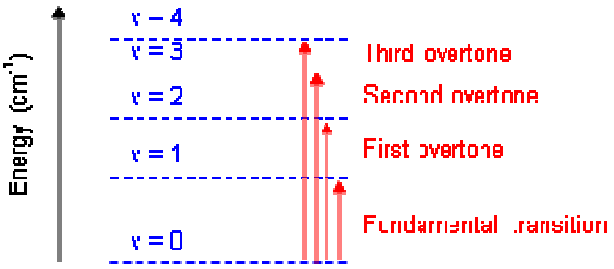


Figure 4. Diagram of the ground electronic state which its vibrational levels, showing the concept of overtones and fundamental absorption

Some things to remember:

- In order for a molecule to absorb infrared light, it has to vibrate at the same frequency (wavelength) as the irradiated light and change its dipole moment
- The absorption of infrared light induces to organic molecules to achieve more energetic levels due to an increase of its vibration: higher vibrational levels.
- Absorption of MIR light induces to fundamental transitions (jump to a consecutive vibrational level) while absorption of NIR light induces to overtones (jumps of more than one vibrational level)
- To achieve higher overtones a molecule need more energetic light: NIR light with shorter wavelengths

1.4) Using NIR Spectra for Analytical Purposes

Now we have seen what happens to the molecules when absorb NIR radiation, we can talk about how to use this phenomena to carry out chemical analysis. Let's imagine we have a sample of soymeal and we want to know how much protein we have using NIRS. We will follow some reasoning steps to figure out how the radiation can be used for this purpose.

1.4.1) Principles of Beer's law

Knowing that NIR can be absorbed by organic compounds, we can think that protein will absorb NIR radiation. What comes next is the application of one of the most important laws for techniques based on radiation absorption: **Beer's law** (Equation 1). This law states that you can correlate the absorbed radiation from a sample with the concentration of a particular compound (C_x).

Equation 1. Absorbance = Cx• e • l

Cx = Concentration of the test solute
e = Molar absorptivity of the test solute
l = Path length traveled by the radiation

The law is directly applied when working with mid-infrared, because the absorbance from a single wavelength can be correlated very well with a certain compound concentration. Which wavelength should be chosen? As previously said, a specific molecule to get to the next energy level needs a specific wavelength. This has been studied for years, so current scientists know which wavelength has to choose and measure the absorbance from knowing the compound of interest. Note how the path length is also included in the equation. Since the molar absorptivity is constant for a same compound, if the pathlength is kept constant and you have several samples of known concentration, you should be able to create a correlation curve, called **calibration** or **regression** line like the one shown in figure 5. Once you have that curve and you determine the equation often by least squares (slope and intercept of the line), you will be

able to predict the unknown concentration in new samples just plugging the absorbance value and leaving the concentration as the unknown.

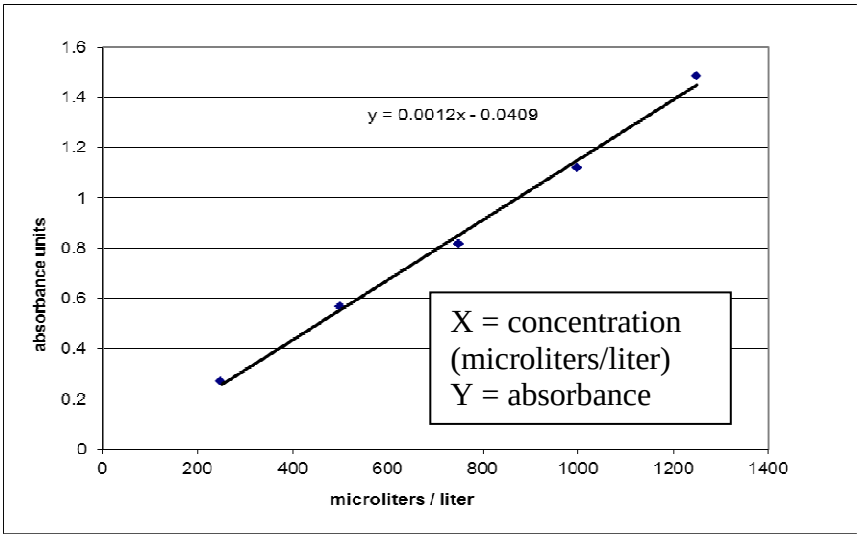


Figure 5. Regression or calibration with MIR spectra, using the absorbance from a single wavelength versus the concentration (microliters/liter of a specific compound) of standards. The displayed equation can be used to predict the concentration of new samples with unknown concentration,

This concept is not so straight-forward when using NIR. NIR absorptions are weaker and some other assumptions of the equation such as no correlation between multiple absorbers, sample homogeneity, or negligible light scattering. Furthermore, it is not possible to use the absorbance from a single NIR wavelength to accurately measure the concentration of a compound. Still, a calibration or regression model is still needed and the law is still true: You can correlate the NIR absorption from a compound with its concentration, but some tricks explained later are needed.

Some things to remember:

- NIRS is based on Beer's law which states that the absorbance of a compound is correlated with its concentration in a sample
- Pathlength is of high relevance and should be kept constant (check Beer's law)
- Beer's law can be applied directly to MIR absorptions, but not to NIRS
- A calibration or regression is needed in order to predict the compound of interest in new samples. That is to say, you have to measure first samples with known concentration of the compound of interest.

1.4.2) Measuring absorption of radiation

When a sample is irradiated with light, according to energy conservation law, fractions are **reflected**, **transmitted**, and **absorbed** (Figure 6). The proportion of each depends on the light wavelength and sample properties (composition and thickness among others). If in order to work with NIRS we need to measure the absorption, how do we measure it? Probably the first answer is "with the right instrument". But, in fact, how can you measure how much light a samples absorbed? Although absorbed light cannot be directly measured, transmittance and diffuse reflectance can be correlated to light absorption according to Equation 2 and 3, respectively. The measurement mode, either transmittance or reflectance, will of course influence instrument characteristics such as the detector position (Figure 7)

Equation 2.

Apparent Absorbance = log(Po/P) = log(100/T(%))

Equation 3.

$$Absorbance = -\log\left(R_{relative}\right) = \log\left(\frac{1}{R_{relative}}\right) = \log\left(\frac{R_{standard}}{R_{sample}}\right) = \log\left(1/R_{sample}\right) + \log\left(R_{standard}\right)$$

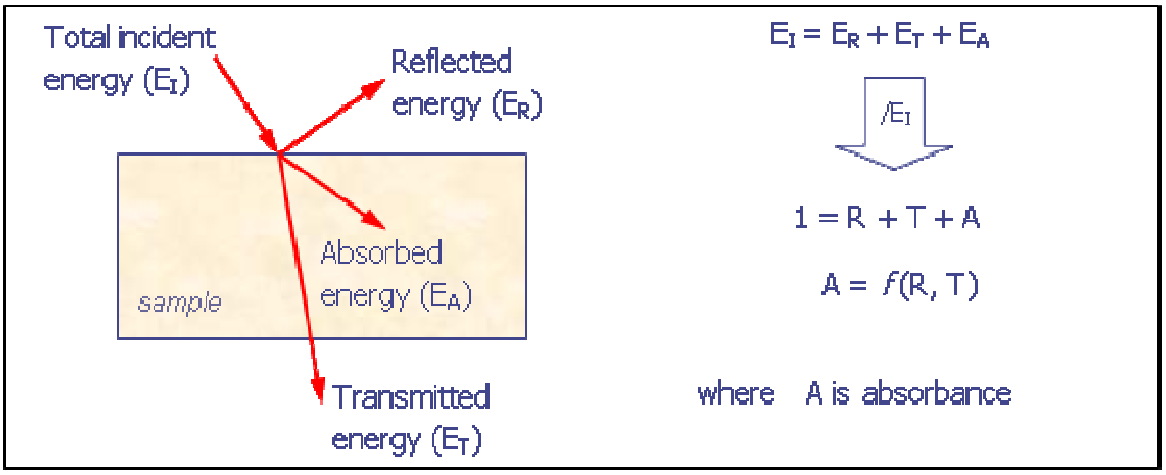


Figure 6. Resulting fractions of incident light when it encounters a sample

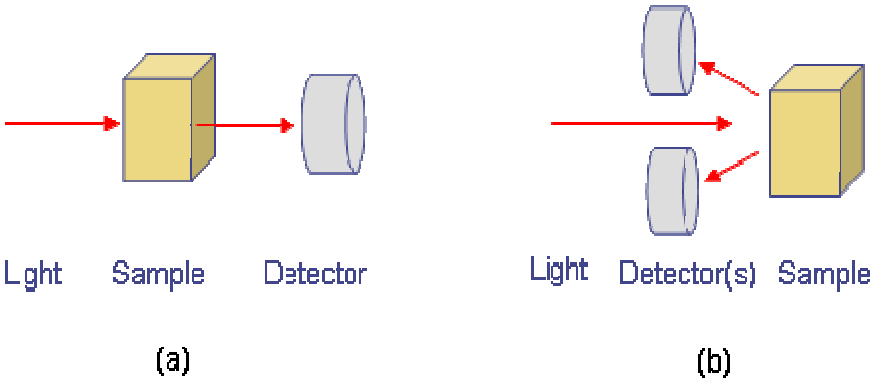


Figure 7. General instrument conformation for transmittance (a) where the detector is placed right after the sample. For reflectance measurements (b) the detector/s is/are placed keeping an specific angle to avoid capturing the specular component

Transmittance (T) is defined as the ratio of radiation passing a sample per unit area (P) divided by the initial radiation power (P/Po), expressed as percentage. That is to say, it is the percentage of radiation which has passed through a sample. Log(1/T), also known as optical density, is called apparent absorbance and it is a close approximation to the real absorbance.

The reflected fraction shows higher complexity. There are two main components of reflected radiation: **specular** and **diffuse**. The specular component angle of reflection is the same as the incident light, is reflected to a single direction, and achieves its maximum intensity when the irradiated light is perpendicular to a smooth sample surface. It lacks NIR relevant information due to its minimum contact with the sample: the sample can't absorb that radiation. The NIR diffuse reflectance component refers to the part of the incident beam that achieves certain degree of sample penetration (few millimeters), it is scattered within the sample while some absorption occurs, and returns to the surface where it can be measured by a detector. Looking at equation 3, Relative reflectance (Rrelative) is measured as the ratio of the sample measured reflectance

(Rsample) over the measurement from a highly reflective material (Rstandard, with reflectance approximately 100%) such as **Teflon** or **Spectralon**.

Transmittance measurements are best taken at lower wavelengths (700 to 1800 nm is common in instruments that measure transmittance) because they are more energetic, have more penetration power and the absorption of those wavelengths is weaker (Because for a molecule to achieve higher overtones is difficult and less probable, the amount of molecules that will experiment it is lower and thus there will be smaller absorption in that region). Measurements by diffuse reflectance are best taken at wavelengths between 1200 nm and 2500 nm. Measuring diffuse reflectance allows working with thicker and denser samples without inducing as much heating as transmission since only a small fraction is penetrated by the radiation. While sample path length is predetermined and must be kept constant for transmittance measurements as we saw in Beer's law, reflectance measurements are not so strict in their pathlenght but it has to be known that it is highly dependent on the wavelength range used in the analysis and sample characteristics such as density or packing, particle size, and material absorption (Berntsson et al., 1998). Physical characteristics affect reflectance measurements especially at higher wavelengths hence any sample changes will create an additional source of variability and noise in the measurements (Norris and Williams, 1984).

Overall, reflectance measurements show lower sensitivity because by diffuse reflectance a smaller sample portion is analyzed (Corti et al., 1999). Its repeatability of measurements is slightly worse which is more noticeable in heterogeneous samples. In specific applications, those limitations may not create significant errors, or may be mitigated by using of a wider range of wavelengths (Kays et al., 2005). Transmittance measurements exceed the accuracy of reflectance measurements in most pharmaceutical measurements, although analytical sensitivity, signal to noise ratio and limit of detection is highly affected by sample position and changes in geometry (Short et al., 2008). Comparison studies in agriculture fields do not lead to a unanimous conclusion regarding superior performance of any of the two measurement modes (Orman et al., 1991; Williams and Sovering, 1993; Borjesson et al., 2007; Xing and Guyer, 2008). Although there is a general preference towards transmittance measurements when small

concentrations need to be measured, the differences with reflectance measurements usually arise from combination of factors such as selected wavelength range, instrument and sample characteristics, data processing/analysis, and sampling procedure (Kays et al., 2005; Short et al., 2005; Delwiche, 1995; Cogdill et al., 2007).

Some things to remember:

- You can not measure a sample absorbance directly.
- You can measure the reflectance and transmittance of NIR light with detectors, and correlate any of those measurements with absorbance
- Transmittance measurements are carried out with shorter wavelengths and require a specific fixed pathlength. When working with transmittance mode be careful with termolabile samples and sample position
- Reflectance measurements are more flexible since allow working with thicker samples, although they are affected by sample physical characteristics. Be aware of sample heterogeneity.

1.4.3 The NIR spectra

The term **spectrum** (its plural is **spectra)** refers to the group of absorbance, transmittance or reflectance measurements carried out at diverse wavelengths. Figure 8 shows an example of NIR spectrum with the different regions. The highest overtone happens at shorter wavelengths. The last region in wavelengths from 1900 nm to 2500 nm approximately, is called the **combination bands** region. That region has a mixture of information resulting from basically involves a combination of vibrations from the same chemical groups of the overtones, but as a result of interactions between molecular vibrational frequencies, overlapped information from Fermi resonances, and inactive MIR bounds among other complex phenomena (Bokobza, 1998).

The X axis of any tipical spectra has the wavelengths in nanometers or wavenumbers (not so common), and on the Y axis the absorbance units which have been calculated through the conversion of transmittance. Probably the correct way to represent it would be through points, but commonly the spectrum is a line drawn joining all the measurement points.

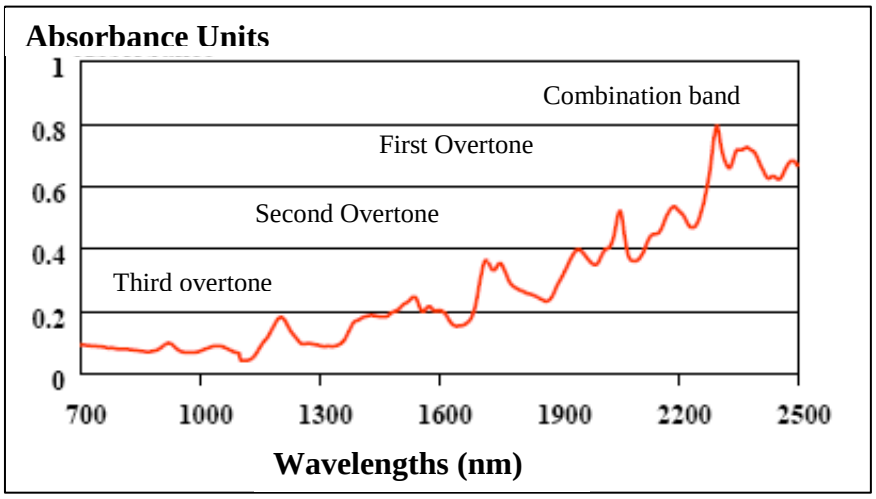


Figure 8. An example of a NIR absorbance spectrum with its parts

Figure 9 shows three different spectra from three different soybean samples. You may not see much difference among those three spectra, but when you use the right mathematical and statistical techniques explained later, you can see that those three spectra belong to three different samples. NIR spectra usually look broad. For instance, compare the NIR absorbance soybean spectra with the MIR transmittance bio-oil spectra in Figure 10. The wavelength units are wavenumbers, and besides that, note how MIR spectra have more noticeable and sharp peaks. Those peaks can be easily assigned to specific compounds (that's the reason Beer's law can be applied more directly), but as you see this is not possible in NIR spectra because there are no noticeable peaks. In fact, there are peaks but they are overlapped so unless you treat it mathematically you cannot differentiate them. NIR spectra from any sample showed broad and overlapped low intensity bands, between 10 and 100 times attenuated compared to the sharper MIR fundamental absorptions (Dryden, 2003).

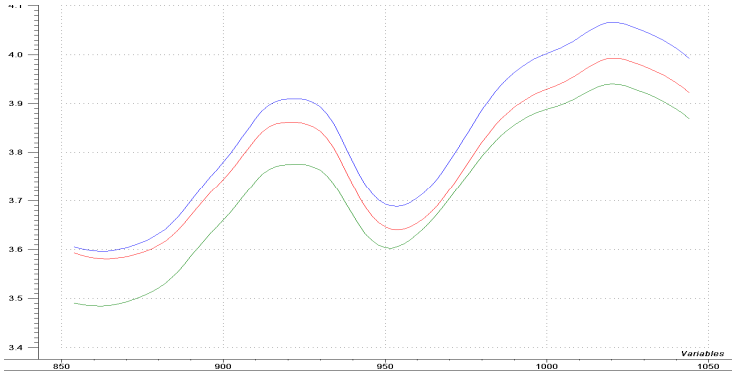


Figure 9. Three NIR spectra from three soybean samples

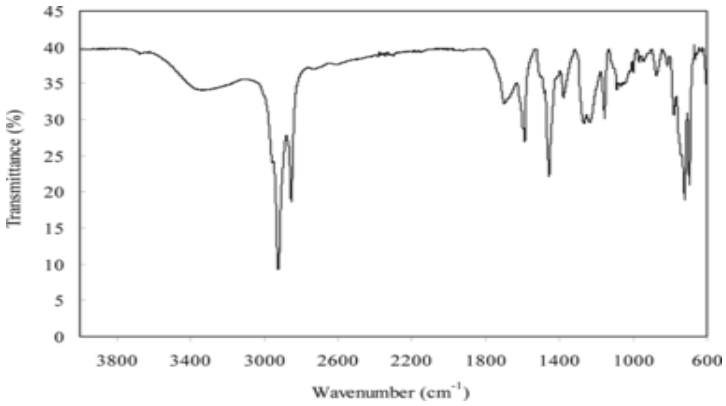


Figure 10. MIR transmittance bio-oil spectrum

Another of the particularities from NIR spectra is the fact that the chemical information is repeated several times. We have seen how it is the result of diverse overtones and the combination bands, so a same molecule can be absorbing the light from the three overtone regions plus absorptions in the combination band region. It is very useful to use the **tables of near infrared absorptions** to guess which wavelengths may be absorbed from a specific compound. Figure 11 shows one example of the table. On the bottom, X axis, you can read the wavelengths and on the top each overtone region is marked. Small blocks indicate the absorption regions of functional groups. For instance, in which wavelengths would water absorb? You would find around wavelengths 950 nm (third overtone), 1430 nm (second overtone), 1930 nm (first overtone) and 2250 nm in combination bands region.

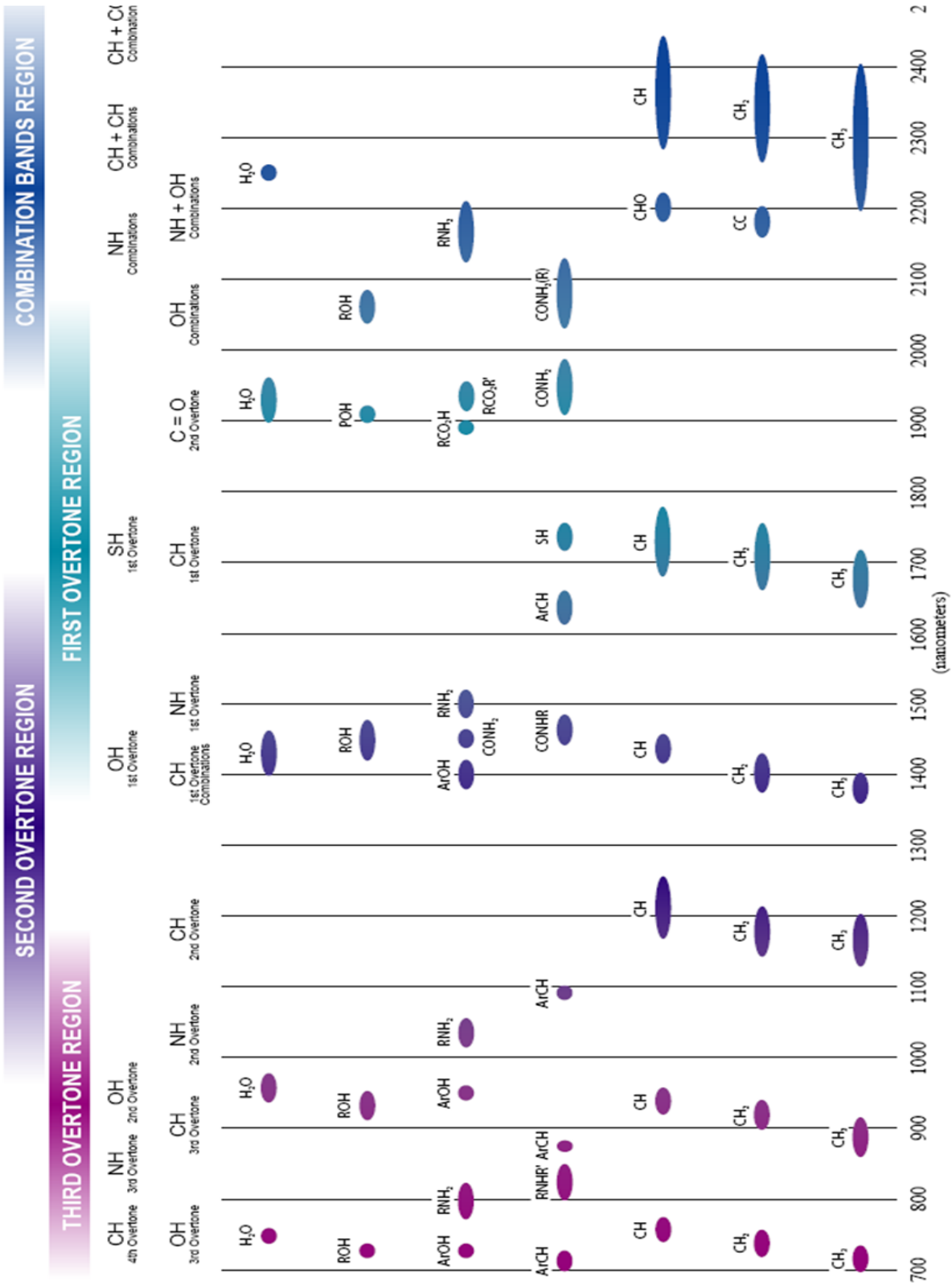


Figure 11. Table of near infrared absorptions

2. HISTORY OF NEAR INFRARED SPECTROSCOPY (NIRS)

The discovering of the near infrared region of the light (NIR) was done by William Herschel (Figure 12), back in the beginning of the nineteenth century. His experiments were based in measuring the heat produced by filtering the sun light on different colors with a thermometer. He realized that temperature increased from going to blue color to red, but amazingly, it kept rising even after positioning the thermometer further from the visible red. His findings (Herschel,1800; also reproduced by Davies (2000)), suggested there were light waves beyond visible light – light that could not be seen. Not further significant research on the NIR region was carried out for 150 years, besides being able to record the absorption of NIR light by organic compounds. This was carried out the first time in 1881 by William D. Abney and Edward R. Festing, two English chemists and astronomers, who using a special emulsion (photographic techniques), measured the spectra of 48 organic liquids. With the advances of spectra-recording instrumentation in the 1950s, few applications using NIR were developed. Most of them were functionality and qualitative studies of phenols, fuels, and polymers (Whetsel, 1968). Researchers found that working with NIR data was not easy: the spectra showed broad bands, and those were not correlated with compound composition with enough accuracy to be solved for measuring compound concentrations. Karl Norris (Figure 13), still known as the father of NIRS, was a United States Department of Agriculture (USDA) employee who became the first to introduce NIRS in agriculture

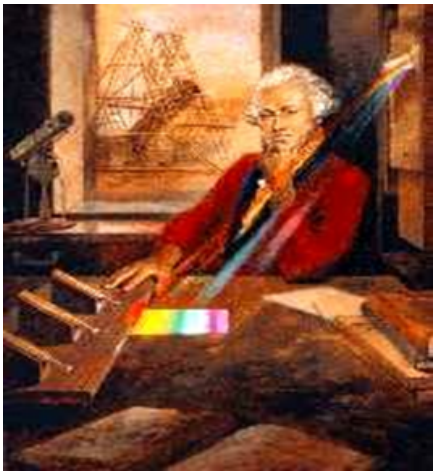


Figure 12. Picture of Herschel, German musician and astronomist, measuring the temperature of light colors

Figure 13. Picture of Karl Norris, called “the father of near Infrared spectroscopy”

and food fields in the 1960s. After his first applications in measuring moisture in whole seeds (Norris and Hart, 1965), he later showed the possibility of carrying out measurements of the NIR reflected light besides the transmitted by the sample after some adjustments to the instrument were made. This lead to big advantages explained later in this manual, such as less sample preparation required. Another of the big propelling factors of NIRS technologies, also attributed to Norris team, was the application of multiple linear regression calibration to correlate multiple absorbance readings at different wavelengths (spectra) with current grain composition. Using the reading from more than one wavelength at a time made possible to finally use NIRS for quantitative purposes. The growth of NIRS the following years was exponential thanks to the gain in power of computation systems, the development of new multivariate calibration and data processing algorithms, and advances in instrumentation.

3. APPLICATIONS

Below there is a list of some of the major practical applications of NIR spectroscopy. Detailed descriptions of most of them could be found in Siesler et al. (2002), Burns and Ciurczak (2001), and Williams and Norris (2001).

Agricultural products and foodstuffs

- Analysis of grain and individual seeds for moisture, protein, oil, starch, and fiber content.
- Measurement of digestion and intake of forages.
- Quality assessment of animal feed.
- Determination of baking quality parameters of flower (protein, moisture, particle size, ash, color, starch damage, and water absorption).
- Determination of protein, fat, and moisture in sliced bread.

- Measurement of sugars and neutral detergent fiber in cereal foods.

- Analysis of milk for fat, protein, lactose, and total solids.

- Determination of the ripening stage of cheese.

- Analysis of alcohol, original gravity, real extract, apparent extract, sugars, total solids in beer, wine, and distilled spirits.

- Measurement of total nitrogen in barley and malt and α-acids in hops.

- Measurement of degrees Brix and acidity in nonalcoholic beverages.

- Determination of degrees Brix, moisture, dry matter, acidity, and firmness in fruits and vegetables.

- On-line monitoring of drying process.

Polymers

- Discrimination between different types of polymers.

- Identity confirmation of various polymer-based fabricated products.

- Discrimination of “in spec” and “out of spec” materials.

- Separation of plastics from nonplastics.

- On-line control of polymerization process in a batch reactor.

- Measurement of crystallinity, viscosity, and particle size/fiber diameter in polymers.

- Determination of homogeneity of blending processes.

Textiles	- Determination of reducing sugars from cotton surface.
	- Control of cotton and polyester blending process.
	- Measurement of the degree of mercerization of fabrics.
	- Determination of cotton fiber maturity.
	- Percent moisture measurement in nylon fibers.
	- Measurement of twist setting of synthetic yarns.
Pharmaceutical and medical sciences	- Identification of individual amino acids (as raw materials for pharmaceutical products).
	- Estimation of active ingredients and water content in tablets.
	- Discrimination between coated and uncoated tablets with identical content of active substance.
	- Distinguishing between types of lactose with different particle size and purity.
	- On-line monitoring of blend uniformity.
	- Measurement of blood substrates (total protein, glucose, total cholesterol, urea, and triglycerides).
	- Monitoring of changes in human blood during storage.
	- Analysis of urine for glucose, urea, creatinine, and protein content.
	- Prediction of fatty acid content in bovine muscle tissue.
	- Analysis of lipoprotein and apolipoprotein composition of the surfaces of living arteries using NIR imaging.

- Differentiation between malignant and benign tissues.

- Noninvasive monitoring of tissue physiology with respect to blood and tissue oxygenation and respiratory status.

Petrochemicals

- Determination of heats of formation, mean molecular weight, and the number of methyl groups per molecule.

- Measurement of octane number, vapor pressure, API gravity, Br number, Pb, S, aromatic compounds, and olefinic compounds.

- Detection of propane and methane gas.

- Determination of methanol in gasoline.

- Determination of moisture and density of crude oil samples.

- Measurement of additives in aqueous cooling lubricant emulsions.

- Determination of concentrations of constituent chemicals in refining operations.

4. INSTRUMENTATION

Despite proprietary instrument conformations, any commercial NIR spectrophotometer has five basic sections: (1) Sample compartment, (2) Light source, (3) Light wave selection system, (4) detector/s, and (5) signal processor or computer. Previously to discuss in detail each section, some popular and relevant terms in instrumentation need to be explain.

Signal-to-Noise Ratio: Signal-to-noise ratio (S/N) is the mean signal level divided by one standard deviation of the fluctuations of the signal. It is popular to think about the strength of relevant signal over the noise. The calculation is performed by instrument companies (it is often given as an instrument specification, the higher the better). Some instrument noise sources include detector dark current, thermal emission of the instrument, non-uniformity of detector arrays, mechanical vibrations in heterogeneous samples, and readout-related noise (Swayze et al., 2003). SNR achievable values in NIR spectroscopy according to Workman range from 25,000:1 to 100,000:1 (Workman, 2005).

Optical resolution: It is commonly measured as the Full Width Half Maximum (FWHM) – the full width of a band at half of its maximum value-. In plain words, it is the smallest wavelength difference distinguished by a spectrometer between two adjacent bands.

4.1) Sample Compartment

Instruments working by reflectance do not need sample confinement for in-line measurements, but it is common to use open sample cups or sample cells confined by silica or quartz (materials transparent to NIR light) in laboratory instrumentation. Transmission instruments may work with confined sample cells as well, but they use specific pre-set pathlenghts ranging from 0.1 to 10 cm, depending on the product to be analyzed (Lipp, 1992). Figure 14 shows some of the common sample cells, the black one is logically for reflectance measurements. The other cells could be used for both reflectance and transmittance, but taking the pathlength into account for the last measurement mode. An integrated adjustable sample compartment with automatic flushing is used for whole grain analyzers. One of the advantages of NIR light is its ability to pass through optical glass fibers preserving most of the signal integrity (losses lower than 5% per km of cable), even if the resulting output intensity is low. This is especially useful for measurements to be made far from the physical instrument and for multiple sampling/ sequential analyses in multiplexer systems. The use of optic fibers with probes for either transmission or diffuse reflectance measurements allows sampling

by immersion in liquids for controlling fermentation or other liquid reaction processes (Buchanan et al., 1988; Tamburini et al., 2003; Sarraguca et al., 2009), contact on small sample areas such as works of art (Bacci et al., 2005), in-vivo medical analysis (Yu et al., 2007), and development of smaller spectrophotometers (Smith, 2000)

Figure 14. Four sample cells from four different instruments

4.2) Light Sources

The most popular NIR light source is the **tungsten halogen lamp**, which has wavelength emission ranges from 320 to 2500nm. The halogen gas allows recycling of the evaporated tungsten (Stark and Luchter, 2005), and brings the advantage of longer lifetime compared to traditional tungsten lamps without halogen.

Light emitting diodes (LED) were used as light source in the first commercial instrument for whole seed analysis in 1985 and in the first portable spectrometers (McClure et al., 2002). Christmass decorative lights are LEDs, although those operate in other wavelength range. The low power consumption, price, small size, and long lifetime (around 25 years) of LEDs still make them the most suitable light sources for miniaturized instruments and specific screening applications outside the laboratory environment (Stark and Luchter, 2005; Axun technologies, 2005). Conventional LEDs

emit in short wavelength ranges (30 – 50 nm) around their center point. Several of them can be mounted in an array with narrowband interference filters if wider wavelength ranges need to be covered, although measuring many wavelengths with this configuration is not an economical approach (Malinen et al., 1998). LED devices have been improved during the recent years to overcome some of their limitations. For instance, some commercial instruments allow easy switching of LEDs according to the application.

Finally, the most innovative light sources are tunable diode lasers, or also called superluminiscent light-emiting diodes (SLED). Using the semiconductor technology of diodes, tunable diode lasers are much smaller than the traditional tunable laser, cheaper, with excellent wavelength resolution, brighter, and with lower noise frequencies than tungsten lamps. SLEDs are suitable for measuring weak absorptions at good signal-to-noise ratio and as light sources in miniature instruments (Lang, 1999). Because they are tunable, this means you can electronically control which wavelength they must emit.

All lights have in common that they need certain time to stabilize their radiation. For this reason, most instruments require warming up time which takes from 20 to 40 minutes.

4.3) Wavelength Selection

Common lamps and even LEDs emit in a continuous range of wavelengths and also detectors read from a wide range as well. As it has been explained previously, any spectrum is obtained having the absorbance (by transmittance or reflectance) values at specific wavelengths. For this reason, there is the need to use any device to select the wavelength of interest from the rest of wavelengths so the detector will measure that single wavelength. Discrete wavelength values can be obtained by filtering the polychromic light beam. The most simple **filters** work by absorption (absorption filters), which are discrete bandpass filters that absorb all light wavelengths but the one of interest. Narrow bandpass interference filters (Fabry-Perot) achieve better spectral resolution and higher output intensity by selecting wavelengths according the refractive index and

thickness of the dielectric material between the two layers of reflective material (Pou Saboya, 2002). To select multiple wavelengths, interference filters are mounted in a wheel which can be automatically controlled to rotate and select the suitable filter for the wavelength selected (Figure 15). This creates spectrometers that provide few spectral measurements. Although filters are an alternative that provides acceptable results, problems of image misalignment and slow operation are common. (Balas, 2009).

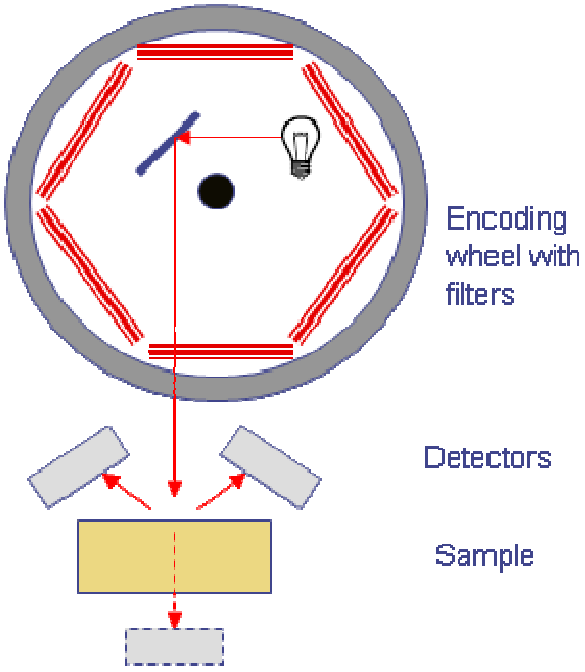


Figure 15. Diagram of a wheel filter instrument

Acousto-optic tunable filters (AOTF) and **liquid crystal tunable filters (LCTF)** allow faster tuning for wavelength selection, and provide better reproducibility without the need of mechanical devices because one filter can select several wavelengths. AOTF filters (figure 16) modulate the light wavelength and intensity through the interaction of sound waves generated in a birefringent TeO2 crystal. The frequency of the acoustic signal makes the refractive properties of the crystal change allowing wavelength-specific transmission. Wavelength discrimination in liquid crystal tunable filters (LCTF) is carried out by applying variable voltage to progressively change the

polarity of a liquid crystal (Garini, 2006). Those filters provide a better output quality compared to AOTF filters, but their short wavelength range is limited (they work below 1800 nm), and give a lower intensity which is dependent on the selected wavelength (Stark and Luchter, 2005; Balas, 2009).

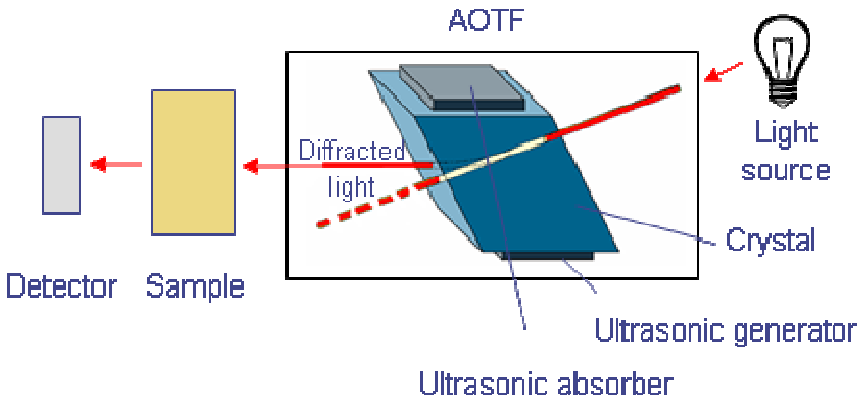


Figure 16. Diagram of an AOTF filter instrument

There are other type of instruments, called **dispersive**, which use a prism or a grating to diffract the incident collimated light beam at different degrees while resolving it in discrete wavelengths. That is to say, when the light hits any of those two elements, it is dispersed according to its wavelengths (the concept is illustrated in figure 17) concept Light dispersion can be done before scanning a sample (predispersive instruments) or after radiating the sample with polychromatic light (postdispersive). Postdispersive instruments offer advantages such as less environmental interferences on the lamp radiation, analyzing wider sample areas, and can hold longer distances between sample and light sources (Schumann and Meyer, 2000; Wang and Paliwal, 2006). Prisms have been lately replaced by gratings because of lower cost and better linear wavelength dispersion of the last ones. A **grating** is a small piece which has grooves or rulings on the side exposed to the light, so when it hits the surface there are phenomena of light interference that leads to the final dispersion. Figure 18 shows two holographic gratings from Shimadzu. There are two types of gratings: Holographic (photosensitive film with

fringes) and ruled (concave surface with fringes). Ruled gratings require being complemented with other optical elements such as lens, and show less stray-light rejection than holographic gratings (Holler et al., 1998; Domanchin and Gilchrist, 2001).

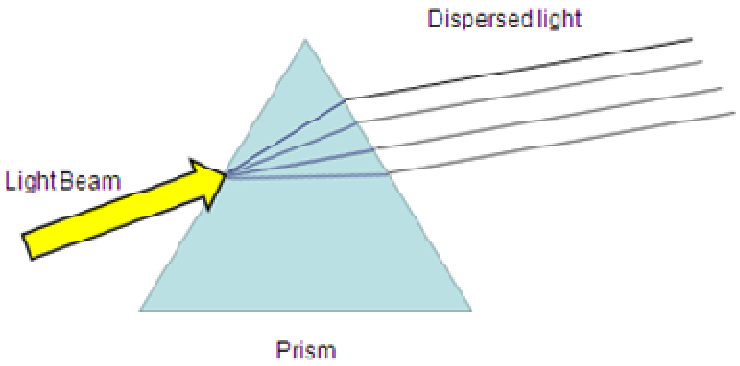


Figure 17. Illustration of the concept of dispersion of light through a prism

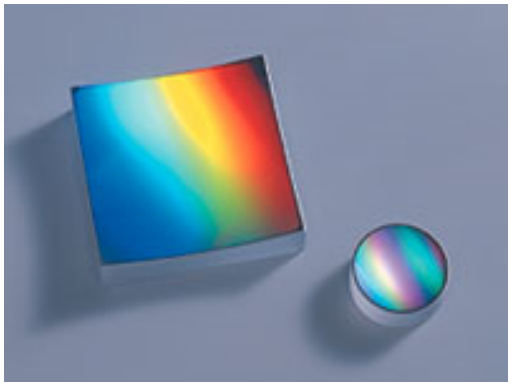


Figure 18. Two holographic gratings
(*http://www.shimadzu.com/products/opt/oh80jt0000001tr4.html*)

In the dispersive instruments group, there are **monochromators** and **spectrographs** such as **diode-array** instruments. Monochromators are pre-dispersive instruments that scan a sample with grating mechanical motion. The basic principle is as follows (Figure 19): Polychromatic NIR light enters through an entrance slit and is then

collimated (light rays are made parallel) by a mirror. The light hits the dispersion grating and later hits a focusing mirror, which reflects it to a second exit slit to either hit the sample (transmittance mode) or hit the single-channel detector (reflectance mode). Entrance and exit slits of a monochromator are very carefully designed to have accurate geometry since they are critical for instrument **observed resolution** (smallest wavelength difference distinguished by the spectrometer) and effective wavelength bandwidth (full width of a band at half of its maximum value, FWHM). When using grating alone without slits, resulting resolution is not enough for most chemical measurements in plastic or pharmaceutical applications (Thermo Fisher Scientific, 2006). Small slits (around 0.1 mm) give low band width, more dispersion, and high spectral definition useful in qualitative applications; large slits (around 2 mm) give more intense radiation and are more suitable for quantitative analysis (Holler et al., 1998).

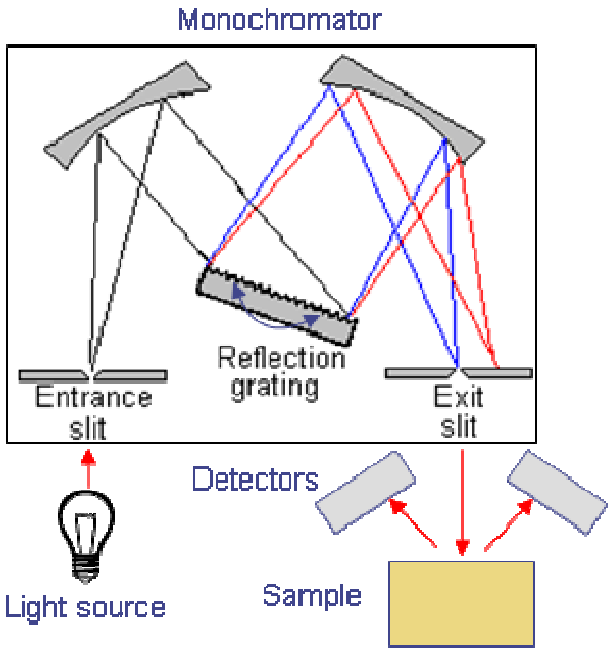


Figure 19. Diagram of a predispersive scanning monochromator with grating, in reflectance mode

Diode arrays spectrographs (Figure 20) are post-dispersive instruments that measure all the wavelengths at the same time thanks to a fixed grating and a set of

detectors placed in array (multichannel detectors). There is no need of exit slits. There are fewer optical elements compared to monochromators and resolution depends on the number of elements in the detector array and array characteristics. The latest advances in wavelength selection besides tunable light sources are the **Micro-Electro-Mechanical Systems (MEMS)** created with semiconductor technologies. MEMS diffraction gratings control light diffraction by electronically controlled movement of diffracting microelements. Their small size and lower cost has led to a new generation of portable instruments.

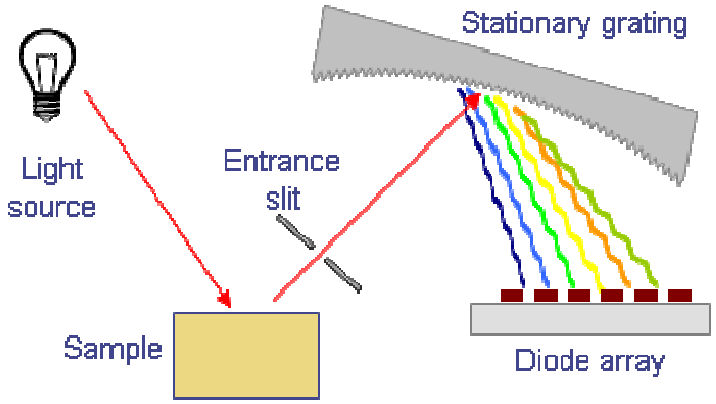


Figure 20. Diagram of a postdispersive diode-array instrument

4.4) Detectors

Detectors transform the incident light energy to electric analog signal. The electrical signal is then amplified and transformed to digital, which may later be further processed by the computer. Detectors and amplifiers are considered the most common sources of non-systematic noise in instruments (random noise). Random noise is reduced in most commercial instrumentation by averaging several spectra from a same sample, improving the signal-to-noise ratio (SNR).

An effective detector must have a linear relationship between the energy input and signal output within its dynamic or working range - from the minimum detectable signal to the maximum before reaching saturation -. Measurement linearity is influenced by other factors besides detector characteristics; for instance, the number of bits of the analog to digital converter device and slight detector misalignments, which can lead to capturing a small fraction of the reflected specular component (often called stray light) in reflectance mode instruments. Without linearity, more complex and potentially unstable mathematics are needed to calibrate the instrument.

Photo-sensitive detector materials are chosen according to the NIR region to be covered. From 400 to 1100 nm, silicon detectors (Si) are common (Stark and Luchter, 2005). Si detectors are stable, fast, not too expensive, and sensitive to low light intensity to achieve good performance. Lead Sulfide (PbS) or Indium Gallium arsenide (InGaAs) detectors can cover higher wavelength regions than Si detectors, being usual having both types combined in a same instrument. **Photodiode Arrays (PDAs)** (Figure 21) spectrographs have a set of InGaAs detectors in array equally spaced or two dimensional **charged coupled devices (CCDs)** (Figure 22). While InGaAs PDAs offer high signal precision, high SNR and less sensitivity to high light intensities when compared to CCD, CCDs have higher signal sensitivity and resolution (Greensill and Walsh, 2000). PDAs take faster measurements (all wavelengths measured at the same time) and can be smaller in size than grating monochromators, which optical conformation cannot be easily reduced in size because it would lead to low throughputs and resolution (Smith, 2000).

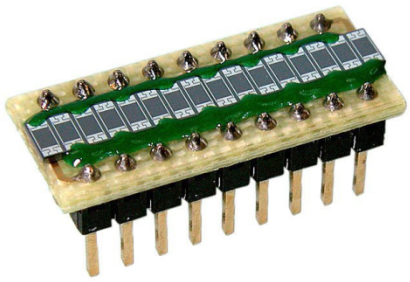


Figure 21. Picture of a Photodiode array (PDA)

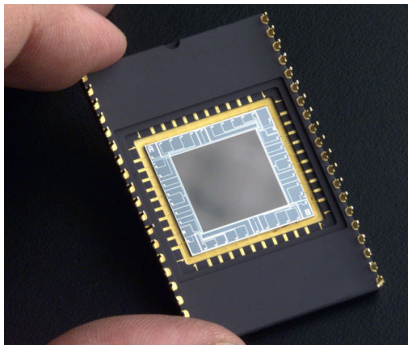


Figure 22. Picture of a Photodiode array (PDA)

4.5) Selecting Instrumentation: General Aspects

There is currently a wide range of instruments with a wide range of prices in the market: small portable instruments for little over 8,000 $, and big sophisticated laboratory instruments over 50,000 $. Instrument price increases with instrument complexity and market position. This need not mean that most expensive spectrophotometers will lead to better performances; indeed, the opposite may be true if no further considerations are taken before purchasing analytical instrumentation. The important point is to know what the instrument function will be and what it could become in future projections. It should be taken in account that calibration costs increases with instrument cost, and the success of any NIR analytical application is highly dependent on data analysis and calibration development up to the point that instrumentation may become a relative afterthought.

To select a suitable instrument, the user must describe the nature of the materials to be tested (sample physical and chemical properties), identify potential applications or uses (environmental conditions and variability in sampling procedures), and determine the accuracy required for the analysis (i.e. screening or demanding quality purposes). Those points should be written down before looking at instruments. Instrument versatility

is a relevant aspect for researchers, and for users whose samples show variable composition and physical characteristics. Sampling speed, although usually not a major limiting factor in NIR spectroscopy, must be considered for in-line analysis and process monitoring. For these last applications, grating monochromators would not be recommended as they take longer scanning times and need regular wavelength standardization due to higher number of mechanical moving parts and PDAs would be more suitable. Instrument robustness is inversely proportional to the number of moving parts and dictates its suitability for rougher environments. Due to the reduced flexibility and versatility of measurements by transmittance in sample presentation and characteristics, in-line monitoring, remote sensing and field applications have been leaded by NIR diffuse reflectance spectroscopy (Drden, 2003; Fischer and Pigorsch, 2000).

Spectral resolution provided by manufacturers, known as observed resolution, affects spectral peak location and hence may impact measurement accuracies. This term can be often confused with wavelength sampling increment also in nanometers (wavelength increment between two consecutive measurements or data points in the spectra), which is greater than resolution. Although high resolution (0.1 nm) may look desirable, it is not always required for success. In analyzing biological or materials with complex composition, resolution shows low impact since the NIR absorption happens over broad regions (Armstrong et al., 2006; Mcarthur and Greensill, 2007). Economical instrumentation with resolution over 4 nm is common and provides acceptable performances in many applications. Resolutions between 1 and 2 nm were required to obtain satisfactory discrimination of compounds with good accuracy when analyzing complex chemical matrices of pharmaceutical and mineral compounds (Mcarthur and Greensill, 2007; Chung et al., 2004). Although both resolution and SNR affect instrument sensitivity and selectivity enhancing SNR compensates for limitations caused by lower resolutions (Greensill and Walsh, 2000; Tarumi et al., 2009, Wash et al., 2000).

Technical support, periodic maintenance and training by the supplier are important. Instrument maintenance is expensive because most operations beyond replacing the light source need to be performed by supplier personnel. Customer services

availability, quality and training are valuable support, especially during the initial stages of instrument set-up, data collection, and calibration development. Several aspects from the data acquisition software have a direct impact on instrument user-friendliness and efficiency on data managing. Current instrumentation has a wide selection of data file and calibration formats requiring varying user knowledge in data handling. Any facility in managing data is highly desirable. From the time spent since the data is collected and the calibration is developed, 80% can be spent on arranging and organizing the data (i.e. exporting, setting the right formats) and just 20% on the real data analysis (Hurburgh and Rippke, 2008).

4.6) List of Popular Manufacturers of NIR Spectrometers

1. ABB, www.abb.com.
2. Analytical Spectral Devices, www.asdi.com.
3. Avantes, www.avantes.com.
4. Axiom Analytical, www.goaxiom.com.
5. Brimrose Corporation of America, www.brimrose.com.
6. Bruins Instruments, www.bruins.de.
7. Bruker Optics, www.brukeroptics.com.
8. BUCHI, www.buchi.com.
9. Carl Zeiss, www.zeiss.com.
10. Control Development, www.controldevelopment.com.
11. DICKEY-john, www.dickey-john.com.

12. FOSS, www.foss.dk.

13. Jasco, www.jasco.co.uk.

14. HORIBA Jobin Yvon, www.jobinyvon.com.

15. Kett, www.kett.com.

16. NIR Technology Australia, www.nirtech.zip.com.au.

17. Ocean Optics, www.oceanoptics.com.

18. PerkinElmer, www.perkinelmer.com.

19. Perten Instruments, www.perten.com.

20. PIKE Technologies, www.piketech.com.

21. Thermo Electron, www.thermo.com.

22. Unity Scientific, www.unityscientific.com.

23. Zeltex, www.zeltex.com.

5. OTHER NIRS-RELATED TECHNOLOGIES

There are other NIRS technologies and instrumentation use of NIR light under slightly different principles from traditional spectroscopy. Two of the most well-established are Fourier transform NIR (FT-NIR) and NIR chemical imaging. Other emerging technologies specifically in medical fields such as NIR fluorescence are not discussed in this review.

5.1) Fourier-Transform Near Infrared Spectroscopy

Fourier Transform (FT) is widely popular in MIR spectroscopy, and it has recently gained high popularity in the NIR range as well. FT technology offers advantages such as high SNR, high light outputs due to absence of slits, fast measurements, instrumental simplicity, and high resolution and accuracy (Thermo Fisher Scientific, 2006). Brimmer et al. (2001) claim that those advantages are more perceptible when working in the MIR region due to the limitation of higher detector noise relative to signal when working in the NIR region.

FT-NIR measurements are carried out in time domain and the direct instrument output from sample scanning is an **interferogram** instead of a spectrum. NIR interferometers (figure 23) split the NIR light beam in two; one of the beams is reflected to a fixed mirror, and the other is reflected to a mirror that moves forward and backward at carefully controlled speed – usually tuned by a HeNe laser-. The reflected beams are recombined back in the beam splitter to generate the interferogram signal, which is a result of light interferences. When displacing the moving mirror, the pathlength difference in relation to the fixed mirror change, leading to different grades of interference between the two reflected beams and which are correlated with different light frequencies. After the interferogram light reaches the sample, transmitted or reflected signal is read by the detector in time sequence (ms), hence measurements are fast. Although interferograms contain information from all the frequencies or wavelengths encoded, it has to be first processed with the Fourier transform. The computation takes as an input a time domain wave signal (the interferogram) from which the transform principle states signal is made from an addition of sinus and cosinus of a set of individual wave frequencies. The processed signal or output looks like the spectra obtained by any traditional spectrometer, but with the expectation of higher throughput and frequency accuracy. One of the drawbacks is the fact that FT-NIR instruments are complex and expensive, and mainly suitable for controlled environments (such as laboratories) due to their sensitivity to external factors such as temperature and vibrations.

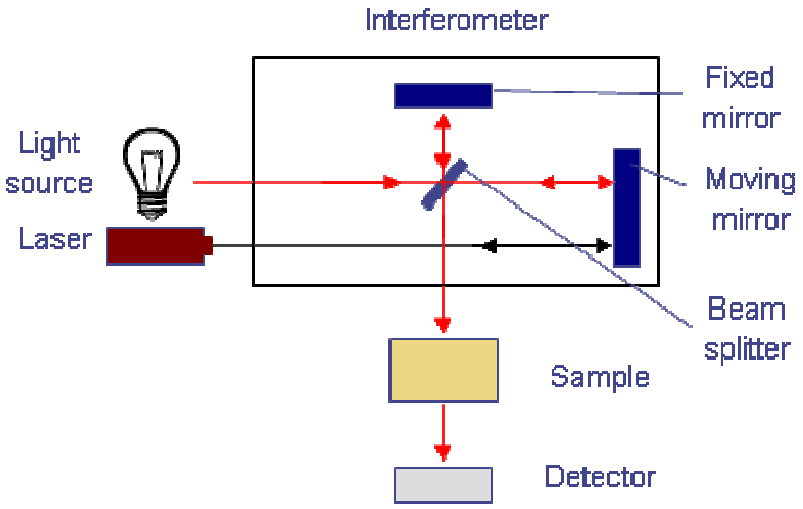


Figure 23. Diagram of a interferometer for Fourier-transform transmittance measurements

5.2) Near Infrared Chemical Imaging

Near Infrared chemical imaging (NIR-CI), or also called NIR hyperspectral imaging, has rapidly become popular, especially in measurements by diffuse reflectance. It combines the advantages of near infrared spectroscopy with digital mapping: the chemical compounds of a sample can be both discriminated and quantified in the sample spatial frame. This is especially useful to analyze compound distribution and sample heterogeneity. Instrument parts and operating principle are very similar to traditional spectrophotometers. The sample scanning procedure can be carried out in two ways: 1) by push-broom or moving imager technique, popular for in-line measurements and sensing, or 2) by fixed staring systems.

Pushbroom instruments measure a spectrum from a whole sample by small consecutive areas or lines while the sample platform is moved and their wavelength selection is usually by dispersion. **Staring systems** scan on still samples, one

wavelength at a time, using either AOTF or CLTF filters. The mapping capability of imaging systems is brought by digital cameras with 2 dimensional arrays of detectors (pixels) such as CCDs that are effective in lower light intensities. Pixel size or area analyzed per pixel range 49 to 1,600 squared microns in commercial instruments, depending on selected magnification. Higher magnification (or smaller sample area captured per pixel) will lead to more detailed spatial analysis and a lower dilution effect of the compound of interest within the sample matrix.

NIR-CI data structure can be thought of as a cube or a stack of cards, where two spatial dimensions are combined with a third dimension corresponding to the chemical information or spectra (wavelengths). Depending on the manufacturer, around 320 x 512 pixels are arranged to capture both sample area and spectra. In that previous example, a total of 320 x 512 = 163,840 data points would be generated for a single wavelength and correlated to small sample portions as a chemical map. If the instrument had 200 sampling wavelengths, the final “image” or data cube would have a total of 320 x 512 x 200 = 32,768,000 data points. Although the amount of data generated is large, visual selection of image areas or pattern recognition techniques help discarding pixels with no relevant information.

This concept is illustrated in figure 24, where each squared surface is like a picture taken at one single wavelength and the small squares within represent pixels. In common imaging terminology, “samples” and “lines” specify the number of columns and rows of pixels; “bands” refer to the discrete number of wavelengths, or following the previous analogy, the number of cards in the stack.

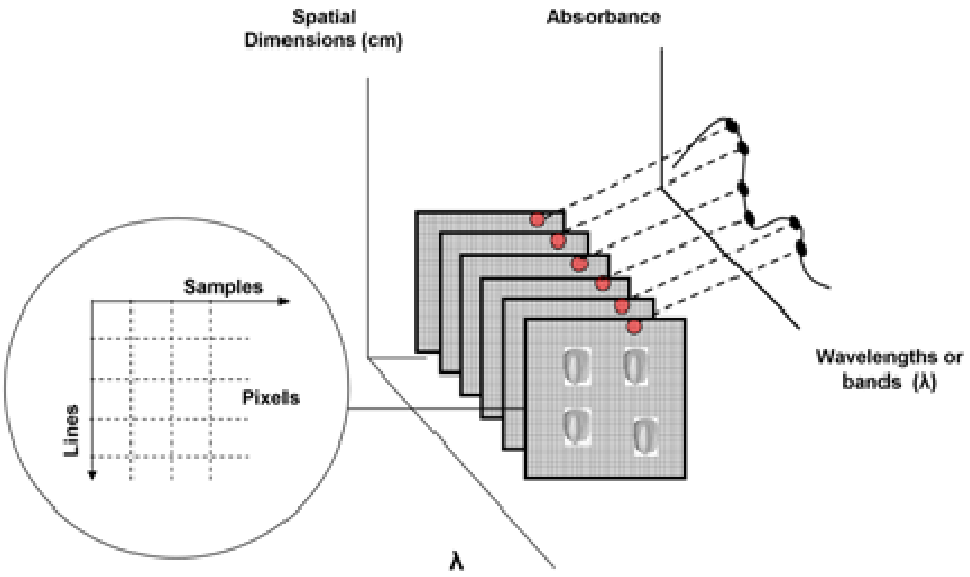


Figure 24. Representation of imaging data cubes in both space and spectral dimensions

6. CHEMOMETRICS AND CALIBRATION PROCESS

As defined by International Society of Chemometrics (ISC), 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. International Union of Pure and Applied Chemistry (IUPAC) define chemometrics as the application of statistics to the analysis of chemical data (from organic, analytical or medicinal chemistry) and design of chemical experiments and simulations. In plain words, chemometrics is the combination of statistics and computers for chemical analysis. Chemometrics made possible the dealing of NIR spectra in resolving highly overlapped and broad peaks, high sensitivity to sample physical characteristics, and the high information redundancy we already explained.

We explained how in order to use NIR information for analytical purposes such as compound quantification through Beer's law, we need first to develop a calibration model. That is to say, sample spectra needs to be correlated with sample compositions by a suitable model which will be later used to predict from unknown sample spectra. Developed calibration models can be loaded to instruments, so when new samples are scanned the spectrometer directly provides the predictions. Discrimination or qualification models stick to the same concept with small variations. Models for predictions are the heart of NIR instrumentation and are part of the American Association of Cereal Chemists (AACC) method 39-00 and the American Oil Chemists Association (AOCS) guidelines Am 1a-09. We will refer to the steps and stages of calibration development in agreement with the methods.

Figure 25 shows a diagram with the basic steps for developing a NIRS calibration. In that procedure, a set of samples are selected and scanned with a NIR instrument. The broad absorptions (spectra) from a sample irradiated with NIR light are correlated with the compound concentration or sample characteristic which user pretend to analyze by a mathematical model. The compound to be measured should either be of organic nature (direct measurement) or be correlated with a sample physical characteristic or another organic compound (indirect measurement). Some relevant aspects of the calibration procedure can be pointed from the diagram on figure 25: 1) there is the need for a fundamental analytical method, called the **reference method**, in order to obtain the dependent variable to be calibrated; that is to say, we need another reliable method which can provide the concentration of our compound 2) a suitable number of samples uniformly covering a wide enough range of analyte concentration should be scanned and be part of the calibration set, and 3) the calibration model should be later **validated** to test the model performance on future samples. Accuracy and precision of predicted concentration values depend on many instrumental (hardware and software) and operator-related factors with the main one being the correctness of the calibration model. Unlike prediction, calibration is usually an expensive and time-consuming process. Therefore, to assure effectiveness of this process, it is very important not only to know the steps involved, but also to understand their implications.

In the case of NIRS instrumentation for grain analysis, calibrations are often preloaded, e.g. wheat protein. Although this may seem an opportunity for new users to save time and resources in developing custom calibrations, the performance of any built-in calibration must be carefully validated to determine its suitability for a particular situation. Calibrations from an instrument brand and model may not perform successfully when loaded to a similar instrument (this will require **standardization** processes, check the advanced topics section), or used on different samples than the original calibration population. Those critical aspects are following explained.

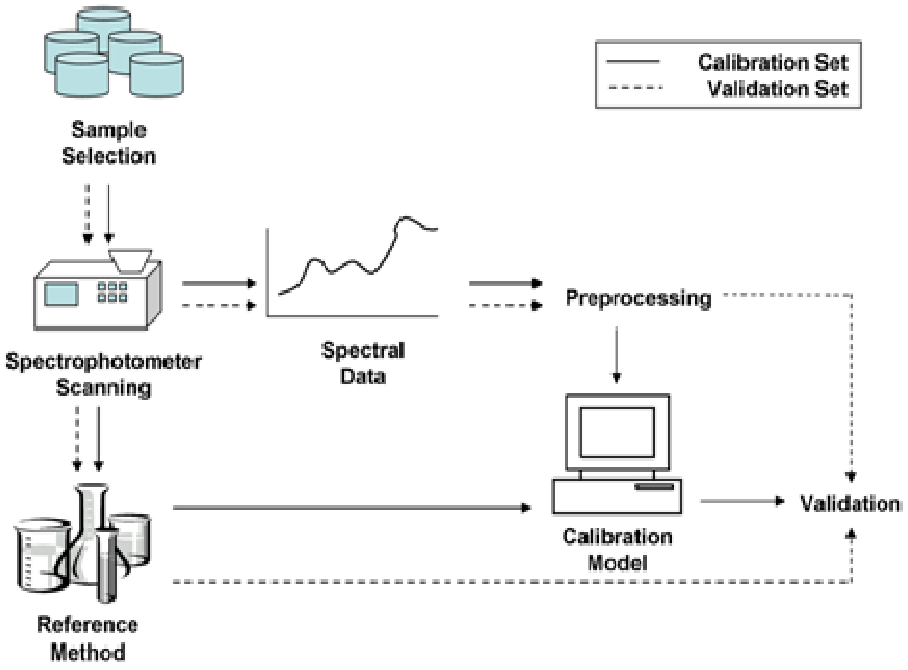


Figure 25. Diagram of the NIR calibration and validation process

6.1) Selecting Samples and Reference Method

The importance of choosing an adequate calibration set is often underestimated and not usually covered in the literature. There is no fixed number or rule-of-thumb to determine the number of samples to be included in a calibration. At least between 20 and 30 samples should be taken for feasibility studies and initial calibrations (Williams, 2001), but more robust calibrations may use few hundred (for instance, instrument built-in calibrations for grain analysis). The number of samples to be taken for calibration also depends on the calibration algorithm to be used. In the AOCS guidelines Am 1a-09 the approximate number of samples is given. Calibrations of homogeneous mixtures (i.e. pharmaceutical powders) may require smaller calibration sets than agriculture samples of high compositional complexity and heterogeneity, such as whole grains or forages.

Users work under the constraints of sample availability and reduced budget to pay for chemical analysis. Nevertheless, there is not enough emphasis on the ultimate consequence of using calibrations developed with inadequate calibration sets: calibrations with low predictive ability. An ideal calibration set should cover the chemical, spectral, and physical characteristics of the population to be analyzed and avoid future extrapolations when predicting new samples (Fearn, 2005). For example, in a case of wheat composition analysis, factors contributing to variation of protein concentration include wheat variety, origin of the samples, their moisture content, sample temperature, etc. Therefore, the calibration/validation sample set for the protein analysis must include samples with different origin, moisture content, and so on. Furthermore, these sources of variation have to be represented equally : The distribution of reference values should be uniform. If the distribution is normal (bell shaped distribution), samples belonging to either higher or lower concentrations have the chance to get more relevance in the calibration, which would not be desirable. Those two concepts can be seen in figure 26.

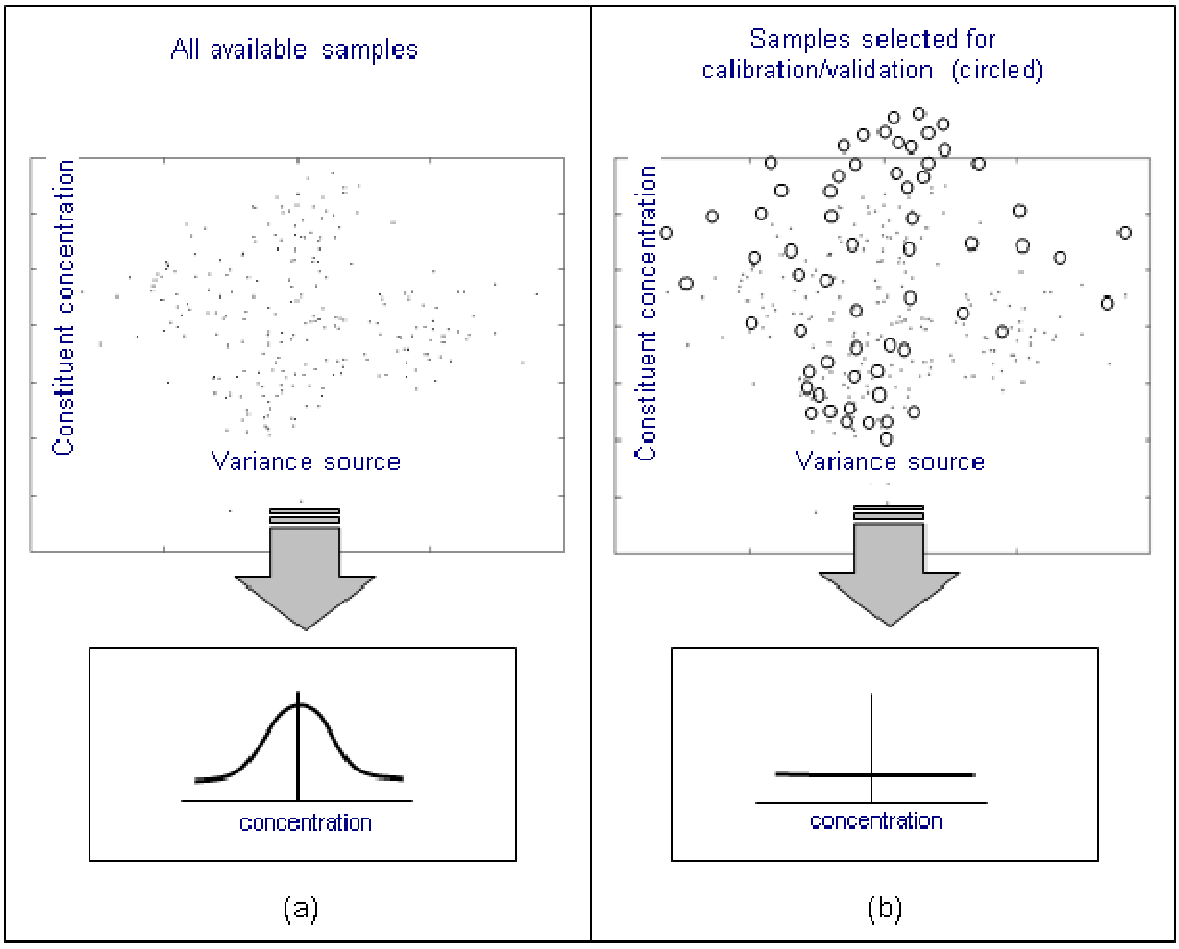


Figure 26. For any compound to be measured, a representative sample of the population usually follows a normal distribution (a). The ideal calibration set should include the entire variance source from the population but follow uniform distribution (b).

Because reference values are not always known and reference analyses of large sample sets may be expensive there are other methods to select an initial calibration set, using spectra. A method developed by Naes (1987) and later illustrated by Naes et al. (Naes et al., 2002c) uses **principal component analysis (PCA)** on the spectra and cluster analysis of the data. PCA is a technique explained later, which basically projects the spectral data to a new reduced dimensional space. Some software have incorporated proprietary algorithms which allow selecting calibration samples based on their spectra or both spectra and reference value, which would be more advisable to obtain robust

calibrations because this way samples are selected taking in account more variability (some unknown sources of variability may be reflected in the spectra).

The chemical reference data

NIRS calibrations can match or virtually achieve better precision (repeatability) and accuracy (closeness to the true value) than traditional wet chemistry methods (Coats, 2002), but paradoxically, NIRS relies on them for calibrations. The quality of the reference data influences NIRS calibrations. In the case of grain analysis, reference methods should follow the AACC and AOCS methods and guidelines. For other products and compounds, an extensive literature research should be carried out to find the best method. Once the method is selected, the selection of the right laboratory is the next critical step. All laboratories, as all measurements, have errors (both random and systematic) but their magnitude and proportions are different. That is the reason why once a laboratory is selected, all samples to be included in the calibration have to be analyzed by the same laboratory. Otherwise, the error from different laboratories would add up, worsening the final calibration model.

One good way to check for the laboratory precision is sending replicates from the same sample, if possible. Equation 4 which is the "Standard Error of the Laboratory" can be used. The smaller the SEL, the better. Laboratory accuracy is impossible to tell unless there is another method which provides known "true values" to compare with. Some programs and associations of laboratories try to bring together different laboratories to agree in the results, and thus improve the overall accuracy of laboratories in the field.

Equation 4. $$SEL = \sqrt{\frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{R}\left(y_{ij} - \bar{y}_j\right)^2 / (R-1)\right]}{N}}$$

With:

y_{ij}: ith replicate and jth sample
$\tilde{y}_j$: mean of the replicates for the jth sample
R: number of replicates
N: number of reference samples

Example. Sending two samples with three replicates, we get the following values (in %) from the lab:

Sample 1: 10.1, 10.5, 10.3
Sample 2: 13.0, 13.1, 13.4

What is the SEL? **Solution: 0.20%**

6.2) Getting the Data Ready

Exporting, organizing, and cleaning the data are the more time consuming tasks of the calibration process. Once the samples have been scanned in a NIR instrument, data is saved in databases following a specific format. Those databases may be such as Microsoft Access files (*mdb extension) or may be a more simple format such as text files (*txt extension). That data usually require to be exported in different format so you can work on the desired software, organizing and cleaning your data. The exporting format depends on the chemometric software the user will use. Virtually all NIR instrumentation manufacturers offer their own software for analysis of NIR data: WinISI and Vision by FOSS, OPUS by Bruker Optics, RESULT by Thermo Electron, CORA by Zeiss, Indico Pro and ViewSpec Pro by ASD, SpectraSuite by Ocean Optics, and so on. Instrument proprietary software is often not very user friendly and lacks of flexibility, more advanced options, use their own terminologies, and the displays are not high quality. However, there are also several instrument-independent (for the most part) nonproprietary NIR processing packages which are popular among users of multiple brands of NIR instruments. Although programs such as R or SAS can be used for calibration development, there are three programs especially known and used in NIRS communities: **The Unscrambler** (by CAMO), **PLS_Toolbox** (by Eigenvector Research Inc.) with **MATLAB**, and **GRAMS Suite** (by Thermo Scientific). Matlab is popular software in engineering fields which require learning programming language, but

offers big flexibility. For this reason, Eigenvector has developed a product for multivariate analysis that does not require the use of Matlab: SOLO (Stand Alone Chemometrics Software). Most of the examples and exercises in this manual have been developed with PLS_Toolbox + Matlab and The Unscrambler.

At the end, the instrument dictate which software to use since every instrument brand only accepts calibration models loaded in specific formats. This is a big inconvenience for laboratories with different instrument brands which end up dealing with different data formats and software. Some common formats you may be able to export from the instrument and which are compatible and importable in most of the software are ascii text files (.txt extension), comma-separated values (.csv), Excel (.xls), and Jcamp – DX. Table 1 is taken from CAMO and shows some common data formats and their extensions which can be imported to The Unscrambler. Files with extension .spc contain spectra information and is generic of GRAMS Suite.

Table 1. Common data formats which can be imported to The Unscrambler

Files:	File Extensions
ASCII	.INP, .DAT, .TXT, and .ASC
Old Unscrambler	.UNS, .UNM, .UNP, and .CLA
Excel	.XLS
Lotus	.WK3 and .WK4
JCAMP	.JDX
APC	.HDR
NSAS	.DA
Tracker	.CAL
Grams	.SPC, .CFL
Matlab	.M
Guided Wave CLASS-PA and SpectrOn	.ASC, .SCN and .AUTOSCAN
Indico	.ASD, .nnnn (any number)
Hitachi F3D	.F3D

Once the data is imported in the selected format, the most common way to organize it is having the wavelength measurements in columns, where each row is a sample. The concentration values (reference values) would go in another independent column. Figure 27 shows how it would look in The Unscrambler interface, in this case there are two compounds (two calibrations can be calibrated).The number of wavelengths depend on the instrument (its wavelength range and its sampling interval; that is to say, the wavelength range between measurements), you may have less than 50 wavelength columns in a filter instrument, but the number of wavelengths or data points (columns) easily reach over 1,000 in reflectance instruments. Note that in terms of statistical analysis the wavelength readings are a big **matrix** ($n_x m$) where *n* is the number of samples and *m* is the number of wavelengths or data points the instruments provides). The reference values from a single compound are placed in a single column, which is a **vector**. This concept is important; remember that NIRS calibrations are based on multivariate analysis (more than one variable, which in this case means several wavelengths), so when talking about calibration models and algorithms it must be used matrix notation.

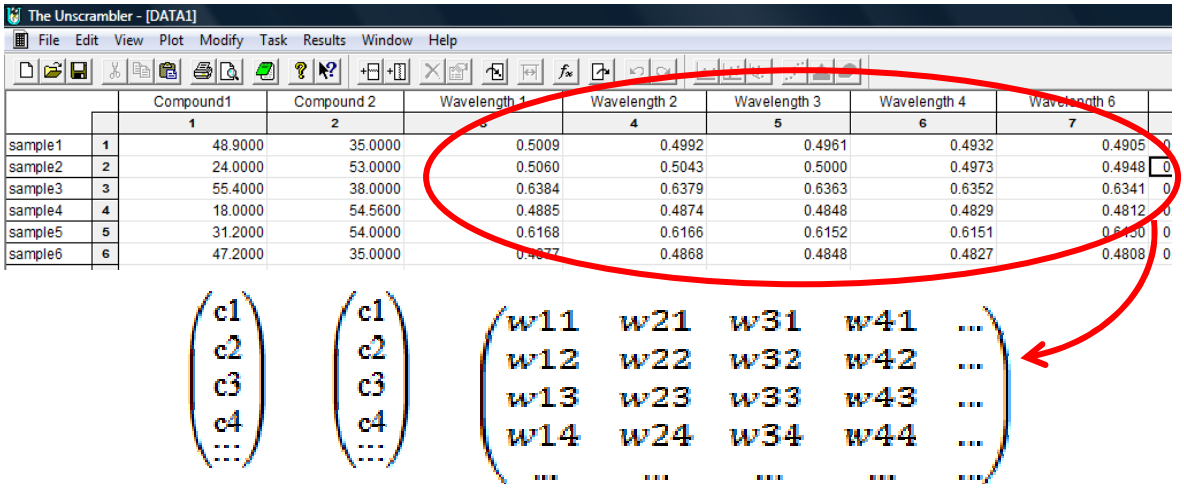


Figure 27. Data organization in The Unscrambler. Each measured compound (i.e. protein, moisture…) would be a vector of concentration values. Those would be the dependent variables (y) in future models. Spectral data can be thought of a big matrix of independent variables (X), which are the wavelengths

Once the data are imported and organized, there should be a check for mistakes and **outliers**. Outliers are those samples that for any reason have higher error associated from either bad reference data or incorrect/problematic sample scanning. Outliers from either reference values or spectral data exist and most calibration methods are highly sensitive to them (Kovalenko et al., 2006; Hubert et al., 2008). Some tests and statistics such as Dixon test (Dixon, 1950) or Grubbs studentized mean deviation (Grubbs, 1950) can be used as an assessment for potential outliers from the reference data a priori. When the problem comes from sample scanning, visual check of the spectra can identify abnormal and noisy spectra as shown in figure 28 with soybean spectra.

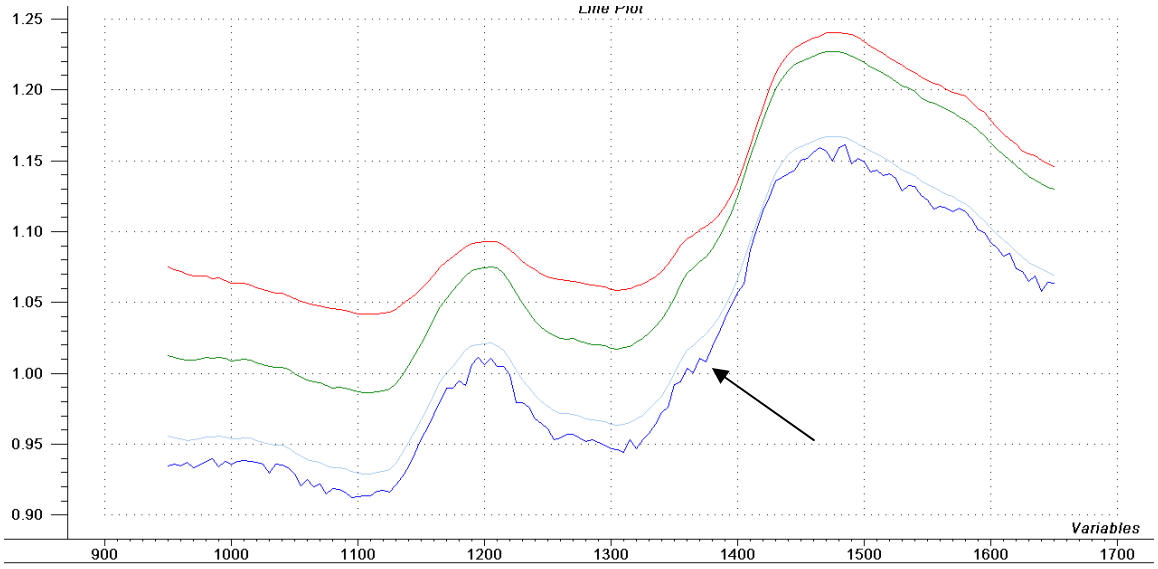


Figure 28. Soybean spectra from four samples. The arrow indicates a noisy spectrum which should be removed from the calibration set.

Visual check is often not enough, and possible outliers may not be detected until data is either preprocessed (next section explains preprocessing methods and its purpose), or a first attempt of calibration has been carried out. Detecting multiple potential outliers is not simple; their effect is masked with each other. Traditional approaches to detect single outliers do not perform well (Walczak and Massart, 1998; Naes et al., 2002a). One

basic approach to detect potential problematic spectra is carrying out **Principal Component Analysis (PCA).** PCA is well known in clustering analysis and data compression because it basically summarizes the information from a high number of variables (in our case, wavelengths) to a set of fewer new variables, which are called **principal components (PCs).** Going a little bit deeper, PCA summarizes the variability from the variance-covariance matrix of the spectral variables, reducing the dimensionality of the data but keeping the main information from the variables. Previously to carry out PCA, data (all wavelength readings) must be normalized by autoscaling or by mean centering (those are preprocessing treatments explained in the following section) so the data will be centered at the new coordinate axis. The data matrix is projected on new **orthogonal axes**: those are the new variables, the principal components (PCs) which are built as linear combinations of the original axes (the original variables: wavelengths). This projection to the new axes is done according to equation 5. Because the axis or new variables are orthogonal, the new data matrix after the projection should not have correlation between variables anymore (remember that NIR spectra shows higher correlation between the wavelengths, carrying out PCA this correlation and redundancy is eliminated)

Equation 5. $T = X*P + \text{residuals}$

X is the original data matrix, where each row are readings from a sample, and each column contains the readings from all samples at certain wavelength (this is the way your data was previously organized). The **loadings** (P) can be understood as the weights for each original wavelength in each principal component. The original wavelengths will contribute differently to the new variables (PCs), for instance one wavelength may be one of the most important in the second PC, and not be contributing significantly in the third. T is the **score matrix**, or the new values that original data acquire on the new dimension. Basically, the new data axes (PCs) will be calculated following the direction of the largest variability of the data. PCs are calculated consecutively: The first PC is calculated following the direction of largest variability; the second PC will be calculated following

the second direction of largest variability but at the same time being orthogonal with the first PC; the third PC will be orthogonal to the first and the second PC and will be calculated following the third direction of highest variability, and so on. The concept is summarized in Figure 29 with a three-dimensional data set (three variables) where two of the variables are correlated (Figgure 29.a). Figure 29.b shows the three principal components drawn from the direction of highest variability (PC1) to the lowest (PC3). In Figure 29.c, data are projected on the new PC axis or new variables, and data dimensionality is reduced to two dimensions after removing the initial collinearity. The number of PCs that can be calculated depends on either the number of initial variables or samples, but commonly only up to 20 are calculated. Only few PCs – the first ones - are considered important at the end, depending on the variance they explain over the total data variance. This can be checked and deduced from the **eigenvalues**. Each PC has an eigenvalue associated (the sum of all the eigenvalues is equal to the number of PCs) which is a constant number obtained through the projection and calculation process of each PC (eigenanalysis). The last PCs will not have important information because they explain very small sources of variability, usually associated with noise. When researchers use PCA to summarize their data, they may use all PCA that have an eigenvalue higher than 1, or commonly, they plot all eigenvalue and do not take further PCs after the first inflexion point of the plot or elbow

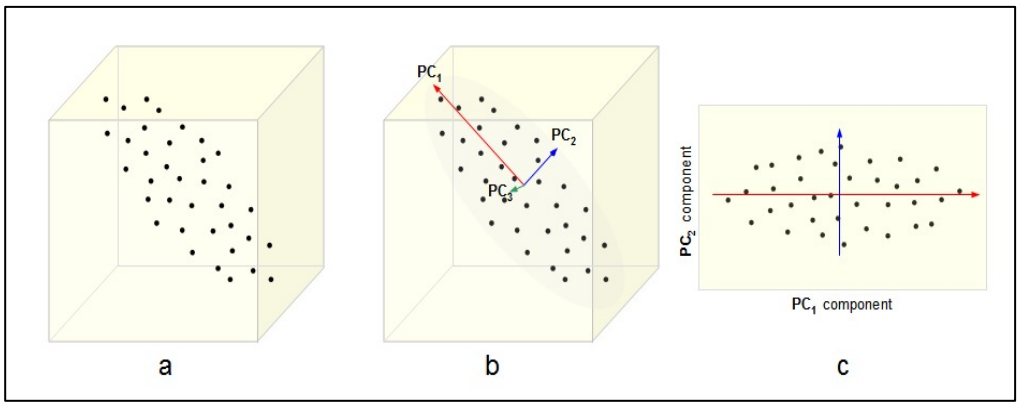


Figure 29. A data set with three dimensions show a clear correlation among two of the variables (a). The three principal components are drawn following the direction of highest variability while maintaining the constraint of orthogonallity among them (b). The third PC can be discarded as it represents small variance (c).

Section 8 has an exercise to carry out PCA which helps clarify these concepts. Available software carries out the calculations for you, so the most important right now is to understand the meaning of the results. Further detail on the mathematical procedure of the whole PCA is out of scope, a big number of useful tutorials can be found on-line. The main concept that user must keep in mind is what PCA does to the data: Summarizing the important information of the spectral data in a smaller set of non-correlated variables called principal components.

So how PCA can help us check for outliers? The fact is that when you plot the spectra, unless a specific spectrum is really abnormal, it is difficult to identify problematic spectra. Carrying out PCA and summarizing the information in fewer variables makes this task easier. Figure 30 shows an example where spectra does not show much unless very close attention (check out the red spectrum), but after carrying out PCA an representing the scores on the second PC and the third, it is easy to see that one sample has some problem in its spectra (Figure 31). This may or may not be an outlier, so it is not recommended to delete the sample but flag it and check how it may affect the final calibration.

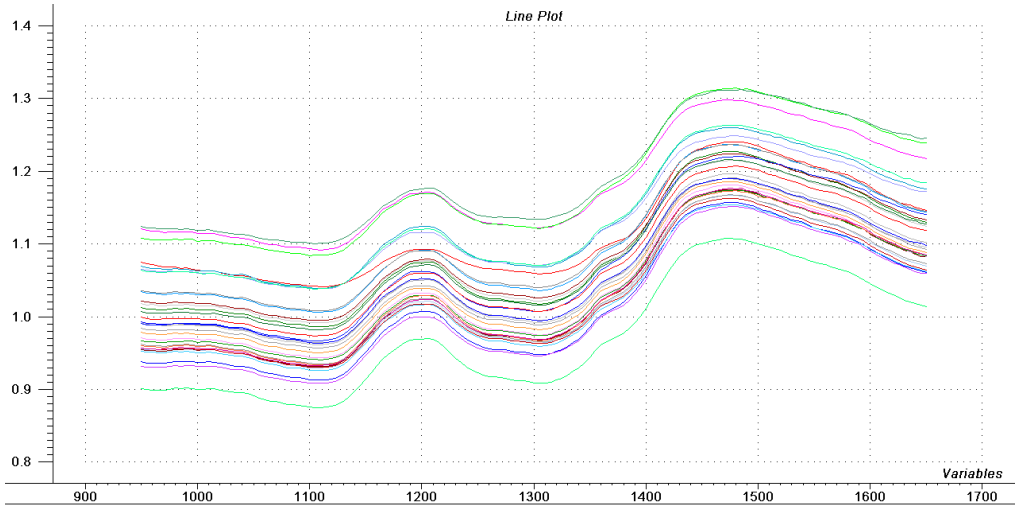


Figure 30. Soybean spectra looking normal at first sight

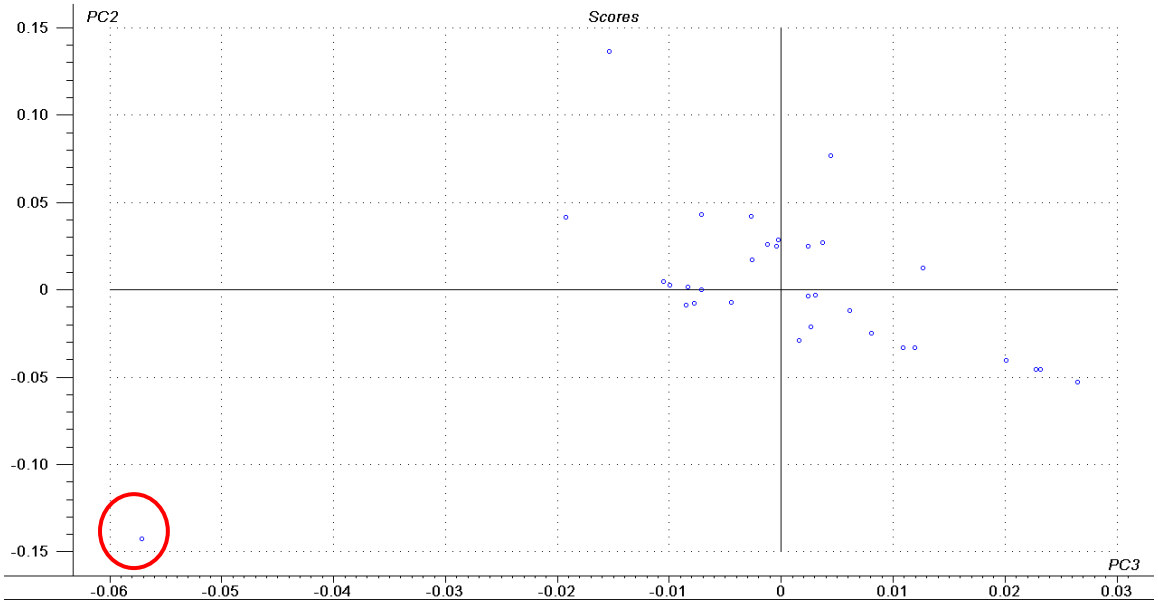


Figure 31. Scores from the second principal component versus the scores of the third principal component showing one of the samples is different from the rest

Measurements of influence such as **Leverage** or **Hotelling's T^2** statistic give an idea of how different a sample is from the rest of the data in a given dimension, as they measure the distance of each sample from the origin. Hotelling's T^2 statistic is distributed as Fisher's distribution but in a multivariate extension. For this reason a sphere of a desired confidence level (a tolerance volume) can be drown in the PCA score plot (PLS_Toolbox has this function in its PCA interface) to help flagging samples significantly different from the rest. Leverage is directly proportional to Hotelling's T^2 and is correlated to the **Mahalanobis distance**, another popular way to measure distances in multivariate analysis (Equation 6, where N is the number of samples an h is the leverage).

Equation 6. Mahalanobis distance = $(N - 1)(h - 1/N)$

Most of the calibration models are based on PCA approaches so both Leverage or Hotelling's T^2 can be used once the calibration model is developed. Plotting sample Leverage or hotelling's T^2 values versus their residuals (what have not been explained by the model, sample error) is a powerful alternative for outlier detection (Naes et al., 2002a). The exexclusion of a sample targeted as outlier from the calibration set could improve the calibration. However, if removed, enough similar samples should remain in the calibration set to avoid significant reduction of representativeness, especially in reduced data sets.

6.3) Data Preprocessing (Pretreatment)

Pretreatments or spectral preprocessing methods are a set of optional mathematical procedures carried out on the spectra before developing a calibration model. Mathematical pretreatment of spectra reduces noise or background information (smoothing techniques) and increases signal from the chemical information

(differentiation). Any pretreatment must lead a robust model with good predictive ability at the end. Basically, preprocessing methods can be classified as baseline correction – normalization, signal enhancement, and statistical filtering of signal noise. Pretreatments can be very helpful but there is always a tradeoff between information loss and noise reduction: when removing scattering effects, the chemical signal may also be reduced.

We have already mentioned one example of preprocessing in the previous section. The conversion of transmittance (or reflectance) values into absorbance is already a preprocessing method which helps making the relationship between sample's optical data and its chemical composition more linear and close to Beer's law. Some of the common data preprocessing methods used in NIR technology are listed below and illustrated in Figure 32. The optimum pretreatment for a given spectra depends on the type of signal (i.e. transmittance, reflectance), sample characteristics, instrument conformation, and application or final goal (calibration or discrimination). There is no absolute or general rule for choosing the adequate preprocessing method; it usually requires a trial-error process guided by experience. Reflectance measurements often benefit from methods that reduce light scattering effects such as MSC or SNV. Sometimes, the predictive ability of a calibration model is not improved with further mathematical treatments. Predictive ability may worsen if preprocessing excessively smoothes the signal, affecting the model ability for predicting new samples (generalization capability). For more details on data preprocessing methods and tips refer to Næs et al., 2002, and Siesler et al., 2002.

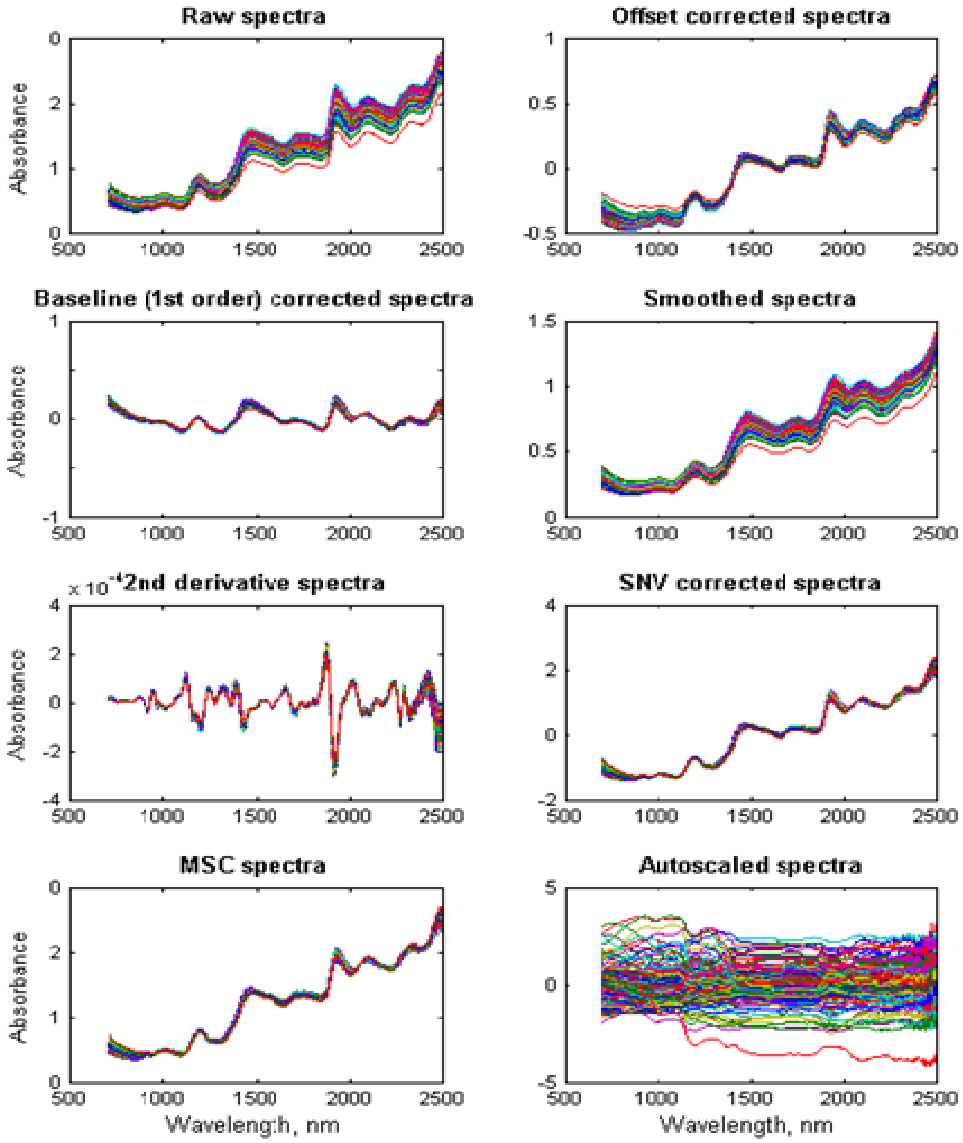


Figure 32. Examples of effects of preprocessing methods on soybean spectra

6.3.1) Offset correction (0-th order baseline)

It removes a constant component from the spectrum. The minimum value of the individual spectrum is removed from each wavelenght:

Equation 7. $$\underline{x_{ik}} = x_{ik} - \min(x_k),$$

where $\underline{x_{ik}}$ is a corrected optical value at *i*th wavelength of *k*th spectrum, x_{ik} is a corresponding raw optical value, and $\bar{x}_k$ is a mean optical value for *k*th spectrum. In figure 32, instead of the minimum value the spectrum average has been removed as set by the PLS_toolbox algorithm, while the first algorithm is adopted by The Unscrambler.

6.3.2)Baseline correction (n-th order baseline)

It removes the constant component and low-frequency noise from spectrum. Check in figure 32 how each spectrum averages zero same as the offset correction but that ascending trend has been removed. The way it is done is selecting two wavelengths of interest and set those to zero; then by linear/quadratic and so on interpolation the rest of values of the variables between them are recalculated. If the original spectra is very step, second or third order baseline correction may work well, but first and second order are more popular.

6.3.3)Savitsky-Golay Smoothing

This is one of the most popular smoothing methods (others are moving average or other filters) which removes high-frequency noise from spectral data. The algorithm fits a specified degree polynomial function to a window points (user-defined) by least squares. So it splits each spectrum in pieces that have the selected number of smoothing points, and then a polynomial fits the data and becomes the preprocessed spectrum. Selecting higher polynomial degrees and small window size leads to high function fit to the data, but the noise is modeled as well, with no smoothing effect. Low polynomial order and wide window size may lead to excessive smoothing and deletion of spectra features containing information.

6.3.4)Savitsky-Golay Derivatives

It includes derivation to the previous smoothing process which differentiates overlapping signal peaks. Although it is possible to work with high degree derivatives,

most of the works in the literature use a maximum of fourth degree for curve sharpening and absorber separation. First and second derivatives are the most common and provide satisfactory results (Hopkins, 2008). First derivative removes baseline offset while second derivative corrects the signal terms that vary linearly across the wavelengths (baseline slope) (Pou Saboya, 2002). Figure 32 shows the effect of a second derivative.

6.3.5)Standard Normal Variate (SNV)

It helps to reduce light scattering effect in spectral data by centering and scaling each spectrum individually, so each has a mean equal to 0 and standard deviation equal to 1:

Equation 8. $$\underline{x_{ik}} = \frac{x_{ik} - \bar{x}_k}{SD_k} \quad ,$$

where $\underline{x_{ik}}$ is a corrected optical value at *i*th wavelength of *k*th spectrum, x_{ik} is a corresponding raw optical value, $\bar{x}_k$ is a mean optical value for *k*th spectrum, and SD_k is a standard deviation of optical values for *k*th spectrum.

6.3.6)Multiplicative scatter correction (MSC)

It reduces light scattering effect in spectral data similarly to SNV (in fact, they are considered to lead to the same calibration results) but MSC is more complex and memory consuming because depends on the whole spectra set.When applying MSC, the spectra is first averaged and each individual spectrum is regressed by partial least squares to the total average. The regression equation slope and intercept represent the additive and multiplicative effects of light scattering, respectively. Finally, each spectrum is corrected for offset (the offset value is subtracted) and each wavelength of the spectrum is divided over the slope. The regression coefficients should be stored and applied to new data.

Equation 9. $$\underline{x_{ik}} = \frac{x_{ik} - a}{b}$$

where $\underline{x_{ik}}$ is a corrected optical value at *i*th wavelength of *k*th spectrum, x_{ik} is a corresponding raw optical value, *a* and *b* are bias and slope coefficients from least squares linear regression of *k*th spectrum vs. average calibration spectrum.

6.3.7)Mean Centering and Autoscaling

Two of the most common normalization methods for variables (spectral or reference). With mean centering, all variables are set to zero mean removing the absolute absorbance value (absolute baseline) and enhancing the absorbance from each individual wavelength. With autoscaling, all variables are set to zero mean and unit variance:

Equation 10. $$\underline{x_{ik}} = \frac{x_{ik} - \bar{x}_i}{SD_i},$$

where $\underline{x_{ik}}$ is a corrected optical value at *i*th wavelength of *k*th spectrum, x_{ik} is a corresponding raw optical value, $\bar{x}_i$ is a mean optical value for *i*th wavelength, and SD_i is a standard deviation of optical values for *i*th wavelength. Autoscaling allows each wavelength to have the same weight or relevance during calibration development. Haaland and Thomas (Haaland and Thomas, 1988) suggest not using scaling when a big part of the spectra do not contain useful information because variables which have more noise than relevant information will get the same importance as the ones with relevant signal. Either one of those pretreatments are **mandatory** for PCA and PCA-based calibration methods detailed in the later section because they reduce the final model complexity, often reducing the number of variables to be employed by one (Haaland and Thomas, 1988)).

6.4) Calibration Methods

Although not excessively time consuming, development of calibration model is the most important and complicated step of the procedure. The objective of this step is to find a relationship between multiple independent variables x_1, x_2, …, x_n (absorbance at corresponding wavelengths) and dependent variable y (constituent concentration). The process of deriving this relationship is usually referred to as multivariate regression. The first assumption when carrying out a calibration is the linear correlation between analyte or the property to be measured and its absorbance according to Beer's law. Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Squares (PLS) are three of the best known calibration methods that work under this assumption. There may be cases where the relationship between sample spectra and reference values is not linear. Any of the previously cited calibration methods can handle small nonlinearities, but when prediction residuals (sample prediction errors) show certain pattern of positive and negative values or plots of predicted versus reference values show appreciable curvature, we are having a clear case of non-linearity (Naes et al., 2002b). This can be seen in figure 33, where the predicted values versus the real values instead of following a straight line show a bow.

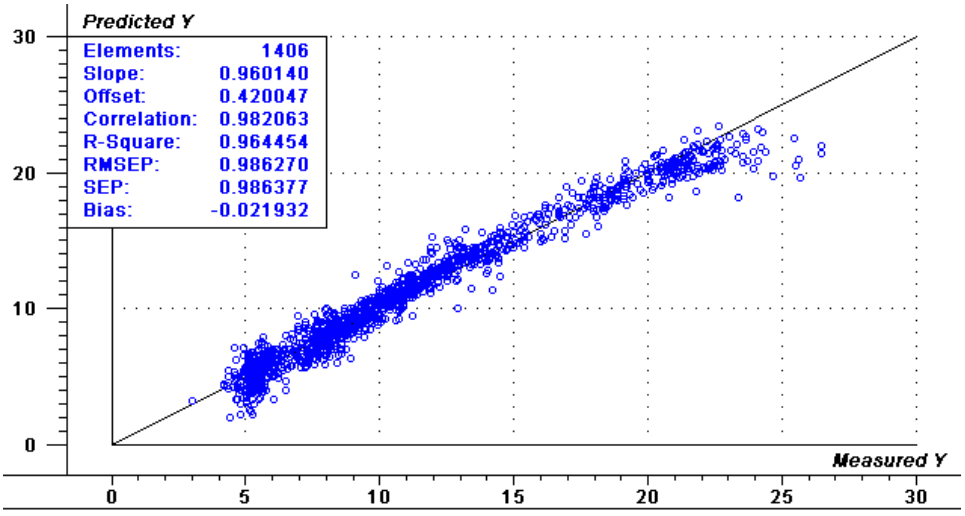


Figure 33. Example of a calibration model which shows high non-linear correlation of the compound to be measured with sample spectra

Often curvature may not be noticeable but for the fact that calibration statistics are not good. Once the problem is detected, there are some solutions suggested by Naes et al. (Naes et al., 2002b) such as trying new preprocessing methods, ignoring wavelengths, adding extra principal components/latent variables to the model, or using non-linear calibration models. Artifical Neural Networks and Support Vector Machines are two complex non-linear regression methods which are not supported by all the instruments. For this reason, linear methods should be tried first.

6.4.1) Multiple Linear Regression (MLR)

MLR is one of the oldest multivariate regression methods that should be used when the following conditions apply:

1) Nonlinearity between optical data and constituent concentration is not detected (or expected). (Note: nonlinearity may be detected, for example, by performing principal component analysis and plotting scores of several principal components versus constituent concentrations.)

2) The number of wavelengths measured is comparatively low (for instance, data from filter instrument) or an advanced wavelength selection method such as Genetic Algorithm could be applied (check the advanced topics section for more information)

3) Independent variables have a well-understood relationship to the response (knowing in advance which wavelengths are of interest)

4) No strong collinearity is present among independent variables: the information from each of the wavelength measurements is not correlated with any of the others.

MLR is a generalization of the univariate inverse method based on least squares fitting of y to $\boldsymbol{x}$. The algorithm gives a linear regression equation of the form:

Equation 11.

$$\hat{y} = \sum_{i=1}^{n} b_i x_i + error$$

Each independent variable (for $i = 1 \ldots n$) x_i is correlated with the dependent variable (the reference value) and its correlation is measured with the coefficient of correlation r (or coefficient of determination r^2). This is done in a stepwise manner through creation of a sequence of multiple linear regression equations. At each step of the sequence, one variable that makes the greatest reduction in the error sum of squares of the sample data (or the one that provides the greatest increase in the F statistic) is added to the regression equation. The process is continued until some stopping criterion is met or all the predictors are processed. In this manner all possible linear regressions on all subsets of the available independent variables are tested. The subset of predictors that produces the lowest standard error is reported. The error term is also known as **residuals**. One of the problems associated with MLR is that it is prone to over-fitting (Davies and Grant, 1987), when a significant amount of irrelevant information (noise) becomes incorporated into the model.

6.4.2) Partial Least Squares (PLS) and Principal Component Regression (PCR)

Both PCR and PLS successfully deal with wavelength correlation. PCR is a direct application of the principal component analysis (PCA) method, and once the spectral data is projected to the new orthogonal non-correlated dimensional axis (PCs) a regression process by least squares is performed between the projected data and the reference values. Wold's introduction of PLS (1975) (Wold, 1975) was an improved alternative to PCR; Both methods carry out regression on data projected to a new dimensional space,

but the new space coordinates created in a process similar to PCA in PLS regression take in account the information from the reference value matrix, and PLS is thus classified as a supervised regression method; the new variables receive the name of latent variables (LVs) instead of principal components (PCs) as the new variables are not exactly the same as PCs. Both methods proved to perform well in situations in the following conditions:

1) Like with MLR, nonlinearity between optical data and constituent concentration is not detected (or expected).

2) The number of wavelengths is large: Because they are both based on PCA, they deal with multiple wavelength summarizing them in a smaller set of smaller variables

3) There is no need for a well-understood relationship between independent variables and the response.

4) Independent variables are characterized by a strong collinearity because of being a PCA-based calibration methods (remember that PCs and LVs are not-correlated variables)

5) The main objective is simply to create a good predictive model, and the effect of each individual wavelength on the response does not need to be explained.

Although both methods provide similar results, PLS become more popular. PLS accuracies may not usually be significantly higher than those of PCR but they are achieved by including fewer latent variables in the final calibration (Naes et al., 1986, 1986; Hammateenejad et al., 2007; Muñiz et al., 2009). PLS is preferred because the algorithm is faster, models have higher precision, and provides more harmonious calibration models (Kalivas and Gemperline, 2006). PLS is based on projecting the initial

variables (wavelengths) from a X matrix (spectra) on a plane formed with a new set of variables (latent variables, similar to PCs) that are orthogonal and linear combination of the initial ones (the wavelengths), but they are furthermore good predictors – in terms of least squares- of the compound to be measured (reference values Y). There are at least two main algorithms to perform PLS calibrations that advanced users may want to check: **NIPALS** (Non-linear Iterative Partial Least Squares) and **SIMPLS**. NIPALS works slower but is told to be more transparent than SIMPLS, which is faster (Wise, no date). For more details on PLS regression refer to Næs et al., 2002, Martens and Næs, 2001. The final calibration model takes the form:

Equation 12. $\hat{y} = f(\boldsymbol{w}, \boldsymbol{l}) = w_0 + w_1 l_1 + w_2 l_2 + w_{(p-1)} l_{(p-1)} + w_p l_p,$

where $\boldsymbol{l}$ is a vector of new independent variables (LVs), and p is their number. The elements of $\boldsymbol{l}$ are determined by searching spectral data space for successive linear combinations of those original predictors that have the greatest covariance between response (y dependent variable or reference) and *x*-variables. Software like The Unscrambler provides the final calibration equation with a similar form to MLR: It provides n coefficients (where n is the number of wavelengths) and an offset value, so in order to predict new samples you only need to multiply each coefficient to its corresponding wavelength reading and add the offset.

PCR and PLS calibrations are only based on a relatively small number of PCs/LVs because similar to what has been explained with PCA, since they are extracted following the direction of maximum data variability, the last PCs/LVs usually involve noise. If an excessive number of variables are included in the calibration, a fraction of noise is also modeled and the calibration becomes too specific to the calibration set. This phenomenon is known as **overfitting** and leads to a reduction of model accuracy in future predictions. There are different approaches to estimate the appropriate number of PCs/LVs to be kept for the calibration – remember that for PCA, where you were not correlation the spectra to any compound concentration, you could check for the eigenvalues; here you have to take in consideration the predictive ability of the model and

check for good predictions-. One of the most employed uses **cross-validation**, later also mentioned as an approximate validation method. The general idea of cross-validation is to keep a single sample (**full-cross validation** or **leave-one-out cross-validation**) or a group of samples (**k-fold cross validation**) apart and develop a calibration with the remaining samples. The remaining samples are then predicted by the developed calibration (validation) and the prediction values are compared with the real reference values to calculate the error. This procedure is consecutively done until all the samples have been predicted once. The error is finally expressed as Predicted Residual Error Sum of Squares (PRESS). In other words, PRESS is the addition of the squared error from each sample when predicted by the model. The PRESS value can be used to select the number of latent variables or principal components in the final model. Chemometric software does this cross-validation procedure using several PCs or LVs and displays the PRESS value graphically so users may visually select the number of PCs/LVs that lead to the first minimum PRESS value from the plot. Figure 34 shows an example of this plot provided by The Unscrambler (the root mean squares of the prediction errors versus the number of PCs used). The best number of PCs to select would be the one that shows the elbow on the plot as then the value does not significantly decrease: Six PCs would be a good choice to start with.

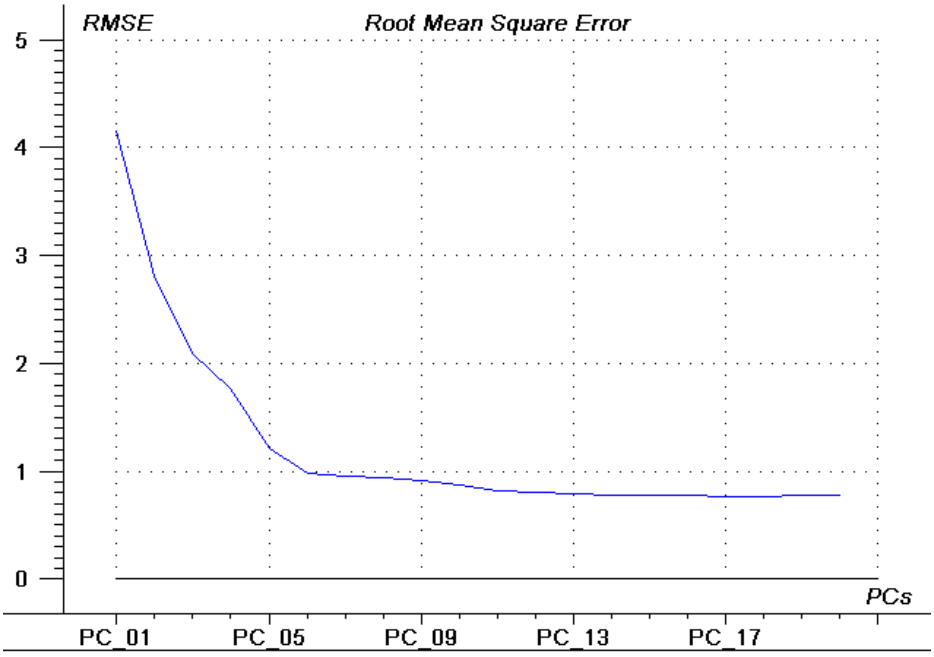


Figure 34. Plot representing the Root Mean Square Error of PLS models using different number of principal components. Users should pick the number of principal components that correspond to the plot elbow.

6.4.3) Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) is a computational method that can be applied to NIR data to develop nonlinear calibrations. By trying to simulate the human nervous system, ANN uses the calibration set to learn about any relationship (no matter how complex and does not need to be linear like in MLR, PCR or PLS) that may exist between spectra and references. ANN regression is much more complex that the previously mentioned methods and require adjusting and optimizing several parameters. Matlab ANN toolbox has very good material to help new users using ANN and has a lot of options to create highly customized nets. The most common type, which we will later describe a little more are the feedforward backpropagation learning nets. ANN regression is best used under the following conditions:

1) Presence of nonlinearity between spectral data and concentration values. The method deals with linear data as well, but because of its complexity and lack of compatibility with most of NIR instrumentation it is best to leave it as a non-linear calibration option.

2) Constituent's concentration of the future samples is expected to be within the concentration range of calibration samples (Extrapolation is not advised in any case, but this is especially true for ANN).

3) The main objective is to develop a model for prediction purposes and there is no need to interprete the wavelength effects on the calibration. ANN is considered as a “black box” because its complexity and the lack of complete information of the learning process that happens in the net. The role of each wavelength in the final model is unknown.

4) This method requires lots of samples to produce robust calibrations. In the AOCS guidelines Am 1a-09 is stated that over 1,000 samples are required for non-controlled environments.

An artificial neural net is composed by neurons (the basic units) or nodes, layers, and transfer functions that join the neurons from different layers (Figure 35). When working with NIR spectra, the **input nodes** would be either the absorbance values from the wavelengths or the scores from principal components, and the **output node** would be the predicted value. Other nodes may be created in **hidden layers** (multilayer perceptron model), which increase model complexity and ability to model non-linear relationships. The nodes are linked by **transfer functions**, which are continuous functions.

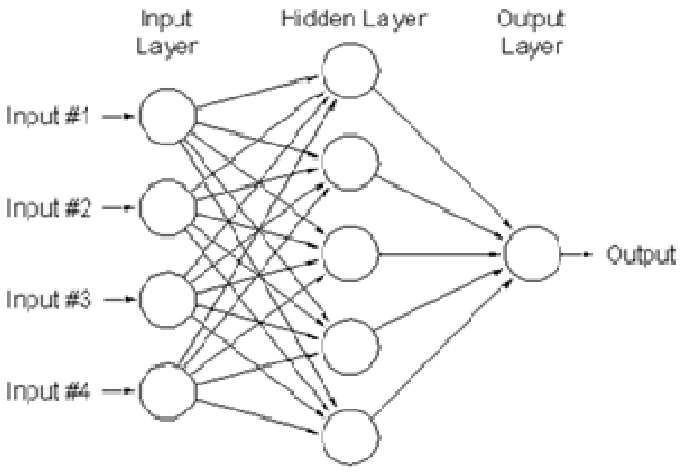


Figure 35. Visual diagram of a neural net with 4 inputs, a hidden layer with 5 neurons, and a single output

Selecting the best net structure for the first training is a matter of experience and a vahue approximation that will be later to be modified for sure. When the net morphology is defined (i.e. number of input nodes and hidden layer nodes), it is trained usually by **feedforward backpropagation** to start the learning process. In training by backpropagation algorithm, random small weights are assigned to each transfer function and are updated according to the prediction error, which is propagated back through the net elements (nodes and transfer functions). This process is done numerous times (**epochs** or iterations) until a minimum established error is met. The learning rate (measurement of the change rate of weights in each epoch) and the number of epochs have to be closely monitored to check for model instability and overfitting (Vandegiste et al., 1998). Using an additional sample set (early stopping set) besides calibration and validation sets is a common practice to avoid the net modeling noise besides relevant information (overfitting).

Once the training is finalized, the final ANN calibration model is a function described by number of hidden layers, number of neurons at each layer (with their transfer functions), and a set of weights (including bias terms) assigned to links connecting the neurons. The equation for a network with D inputs, K neurons in one hidden layer, and transfer functions $\sigma 1$ (output layer) and $\sigma 2$ (hidden layer) shown in takes the form:

Equation 13.

$$\hat{y} = \sigma_1\left[\sum_{j=1}^{K}\sigma_2\left(\sum_{i=1}^{D} w_{ij}x_i + b_j\right)v_j + b_0\right]$$

where X_i is ith input variable, w_{ij} is the weight of the connection from ith input to jth neuron of the hidden layer (number of w-weights is equal to D for each hidden layer neuron); v_j is the weight of the connection from jth neuron of the hidden layer to output neuron (number of v-weights is equal to K); b_j is bias of jth neuron of the hidden layer; b_0 is bias of the output neuron; σ_1 and σ_2 are functions defined, for example, as

Equation 14.

$$\sigma_1(z) = \sigma_2(z) = \frac{1}{1+\exp(-z)}.$$

The dependence of the weights and the amount of parameters involved make the interpretation of each weight and the training process rather complex. Complexity increases as more nodes and transfer functions are added, and it also increases the need for more calibration samples. Another problem that may be encountered associated with the nature of the error function involved in the training process is the high risk of fall in local minima solutions (Despagne and Massart, 1998). That is to say, to achieve a solution with small error but not the best solution overall.

6.4.4) Support Vector Machines (SVM)

SVM regression is the newest method out of the four discussed calibration methods which supposes a more robust alternative to ANN. SVM regression may be a good choice when:

1) Nonlinearity between spectral data and concentration values is present. However, SVM also performs well on linear data, but again it is better to use PCR or PLS in that case because of instrument support and model complexity.

2) The number of calibration samples does not exceed a few thousands. SVM is a computationally intense algorithm, and although it has fewer parameters that need to be optimized, current optimization methods are very memory consuming.

3) Constituent's concentration of the future samples is expected to be within the concentration range of calibration samples; again, extrapolation should be avoided.

4) Similarly to ANN, there should be no need for interpretation of individual wavelengths on the model because it is not possible.

2. The SVM method is based on principles of statistical learning theory developed by Vapnik and Lerner (Vapnik and Lerner, 1963). Initially, the method was intended for solving classification problems, but then was adapted for linear and nonlinear function estimation (Drucker et al., 1997). SVM creates a tube-shaped regression volume with variable diameter. The called “kernel trick” originally introduced by Aizerman et al. (Aizerman et al., 1964) made the algorithm very popular because it opened the opportunity of applying the linear regression algorithm in higher dimensional data: dimensionality does not matter in the final optimum SVM regression function. This basically says that while a linear correlation may not be possible in the initial dimension, the correlation may be linear in another highly dimensional combination of features. The initial data can be mapped to the higher dimensional space applying a mapping function called **kernel** or **kernel function** $\varphi(\boldsymbol{x})$. There are several kernels with variable complexity (polynomial, Gaussian or Radial Basis Function (RBF)…) with which users can experiment, although more complex kernels may be prone to overfitting issues (Ivancicuc, 2007). The final regression model using Least Squares SVM regression (LS-SVM) by Suykens et al., 2002 leads to the following equation:

Equation 15. $$\hat{y} = \sum_{k=1}^{N} \alpha_k K(\boldsymbol{x}, \boldsymbol{x}_k) + b$$

where vector $\boldsymbol{x}$ represents new sample, $\boldsymbol{x_k}$ is kth training sample, α_k is Lagrangian multiplier for kth training sample, b is bias term, N is number of training samples, $K(\boldsymbol{x}, \boldsymbol{x_k})$ is a kernel function defined as

Equation 16. $$K(\boldsymbol{x}, \boldsymbol{x_k}) = \varphi(\boldsymbol{x})'\varphi(\boldsymbol{x_k})$$

LS-SVM model contains information about relevance of each training sample for calculation of $\hat{y}$ and makes its predictions based on relative comparison of new (unknown) sample spectra to the spectra of training samples. Few parameters such as function regularization and kernel parameters (for instance, width in the case of RBF kernel) need to be chosen for optimum prediction ability. However, the number of parameters to be adjusted is much smaller than ANN. Other the already implied advantages of SVM over ANN are fewer samples required and resistance to local minima since SVM uses a **Lagrangian function** that has a single general minimum (Zomer, 2004). More information on SVM may be found in Vapnik et al., 1997, Smola and Scholkopf, 1998, Suykens et al., 2002, and Cogdill and Dardenne, 2004.

6.4) Validation

An adequate validation of the calibration models is a crucial step to determine the suitability of the model to predict new samples, which is the whole purpose of developing NIR calibrations. Ideally, the best validation should be done with distributed samples which were not previously used for calibrating. Ideally, the calibration and validation sample subsets should not be correlated; they have to be assembled independently. However, if the pool of data available for calibration is relatively large (an order of a few hundred samples or more), and it was accumulated over an extended period of time, this requirement does not need to be so strict. Assuming that this is the case, one can proceed

to dividing samples into two subsets. A simple procedure that splits the set into two with the ratio of 3:1 (calibration: validation) can be performed in the following steps:

1) List samples from lowest to highest concentration of the constituent of interest.
2) Starting from the top of the list, transfer three data points into calibration sample set.
3) Transfer the next data point into validation sample set.
4) Repeat steps (2) and (3) for the rest of the samples.

Since independent validation may not always be possible, cross-validation discussed before can provide a basic assessment regarding calibration performance. Be aware that the final calibration model is not tested, but rather several submodels developed with calibration data subsets. That is the reason why any statistic reported from cross-validation cannot be directly compared or interpreted the same way that statistics from a real validation of the final model with new samples. The standard errors from cross-validation are often optimistic and, especially in k-fold validation, highly affected by data artifacts (Naes et al, 2002d). However, reporting cross-validation statistics are preferred over reporting calibration results alone. Table 2 shows the most used NIR validation statistics among the suggested and detailed in Williams (2001). However, it is not unusual to find literature using other statistics, reporting not so relevant figures of merit, or simply not reporting enough information for a good statistical assessment of the model quality.

The coefficient of determination (R^2), which provides an estimation of how much variance between reference and predicted values is explained versus the total variance, seems to be one of the erroneously preferred guides for validation assessment. It ranges from 0 to 1, the higher the better. The calibration model is considered usable for quality assurance applications if r2 is equal to or higher than 0.92 (Williams, 2001). This statistic is highly dependant on the reference value range (Fearn, 2002), so you cannot use R^2 to compare two calibrations unless they have the same range of reference data. It

is common to report this statistic – it has been even abused-, but it should not be reported alone.

The standard error of prediction (**SEP**, or **SECV** when reporting cross-validation results) provides information regarding calibration precision. That is to say, the error you would obtain running the same sample more than once. SEP is corrected for the bias value (or systematic error); thus, when reporting SEP bias must be reported as well. According to a common in NIR community rule of thumb, *SEP* is considered acceptable (sufficiently small or comparable to the error of the reference method) if it is smaller than 1.5 to 2 times the standard error of lab (*SEL*), previously introduced.

The square root of mean standard error of prediction (RMSEP) is related to SEP and Bias according to (Equation 17). Because RMSEP accounts for bias and provides information regarding calibration accuracy, it can be reported alone, especially when bias is small (then RMSEP ~ SEP) (Davies and Fearn, 2006).

Equation 17. $$RMSEP^2 = SEP^2 + Bias^2$$

The ratio of performance of deviation or relative predictive determinant (RPD) is dimensionless and specific of NIRS communities. It is related with the ability of the model to predict future data in relation to the initial variability of the calibration data. Basically, if a calibration leads to a low SEP but the calibration was carried out with a small range of reference values (standard deviation of reference values almost the same as SEP), the model would only be predicting the data average. Williams (2001) provides ranges of RPD values related to the calibration suitability: values above 8 indicate that the calibration can be used for any purpose, while values below 2.3 indicate a poor calibration performance, with use for predicting new samples not advisable.

Table 2. Common statistics used to report the predictive performance of NIR models

Statistic	Units	Equation
Coefficient of Determination (r^2)	Unitless	$r^2 = \frac{\left(\sum_{i=1}^{n} \hat{y}_i y - \sum_{i=1}^{n} \hat{y}_i \sum_{i=1}^{n} y_i \Big/ n\right)^2}{\left(\sum_{i=1}^{n} \hat{y}_i^2 - \left(\sum_{i=1}^{n} \hat{y}_i\right)^2 \Big/ n\right)\left(\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n\right)}$
Standard Error of Prediction (SEP)	Same as reference values	$SEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i - bias)^2}{n-1}}$
Root mean square of the error of prediction (RMSEP)	Same as reference values	$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$
Bias (d)	Same as reference values	$d = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)}{n}$
Ratio of performance of deviation (RPD)	Unitless	$RPD = \frac{Sd_y}{SEP}$

$\hat{y}_i$= i^{th} validation sample predicted value
y_i= i^{th} validation sample reference value
n=number of samples in validation set
Sd_y = standard deviation of reference values from the validation set

Example. The following figures (36 and 37) show the regression lines from two calibrations from a same compound. Which one is better?

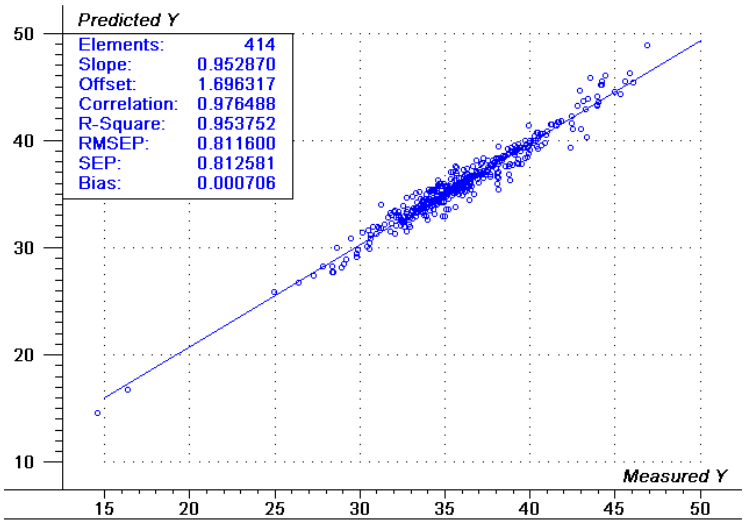


Figure 36. Example 1 of statistics from model validation

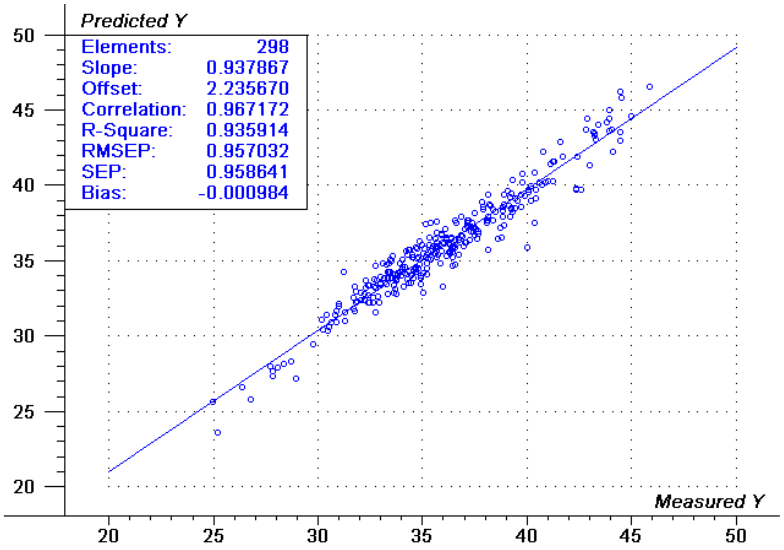


Figure 37. Example 2 of statistics from model validation

Answer. Note checking the square with the software results how the first has higher R^2, lower SEP, and lower Bias (you must check Bias as absolute value, not taking in account the positive or negative sign). So the first calibration looks slightly better.

Note also that the first calibration has two samples with lower concentration which may be very relevant, although we would prefer to have the data evenly distributed through the range, at least having those two samples with low concentration would allow predicting samples with lower concentration more safely – extrapolation, which is prediction outside of the calibration range, must be avoided-.

7. ADVANCED TOPICS

7.1) Optimization Methods

There are developed applications based on the selection of a few variables from a spectrum. Sometimes, certain measurements benefit of using just few wavelengths instead of the whole range. For instance, Foss Analytical patented the use of four wavenumbers (1700, 1407, 1365, and 1238 cm^{-1} or 5882, 7107, 7326, and 8077 nm) to develop models for the prediction of acetone in milk (US Patent 6385549 – May 2002). Variable selection is a type of optimization fairly common in NIRS. Another type of optimization is to use only the most appropriate samples in calibration, the samples that will provide the best fit for the new sample. These models are called local models. Both optimization methods aim to reduce the complexity of the model and can be used together. In this section we expose some methods to both optimize variable and sample selection.

7.1.1) Variable selection

7.1.1.1 Exhaustive search

This is the most intuitive way to select the variables. Basically, is saying “try each variable and combination to get the best result”. The problem of this approach is the time to perform the calculations. For an instrument collecting *n* wavelengths, the total number

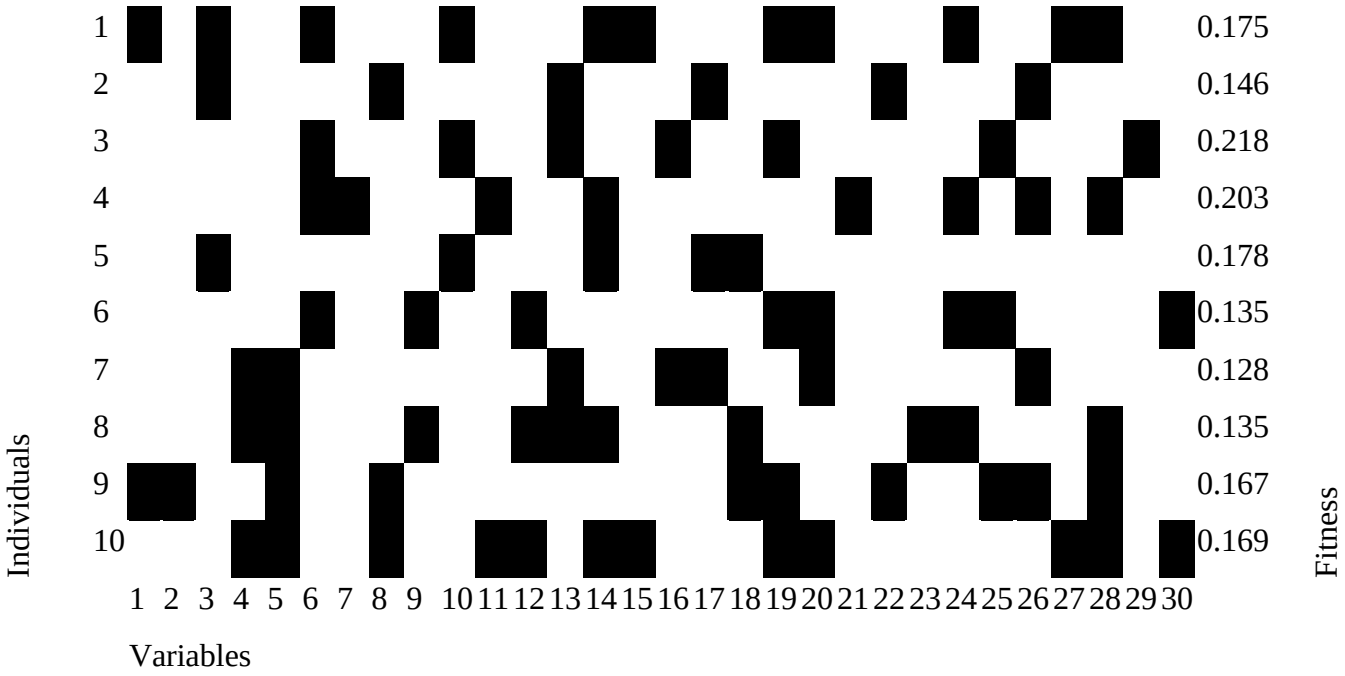
of possible combinations is $2^n - 1$. For n = 100, the total number of possible variable combination is 2^{100} -1 or 1.2677 x 10^{30}. This number is large and, even with a powerful computer, the exhaustive search of the best variable combination takes very long to perform. A type of exhaustive search is **Interval PLS** (Norgaard et al, 2000). Instead of selecting a single variable at a time, a group of adjacent variables are selected; they are called intervals. PLS models based on these intervals are developed and their RMSECV is calculated. The interval presenting the best fitness value is selected. This interval is then combined to other intervals, one at a time. New PLS models are developed and the best combination of intervals is kept. The operation is carried on until the fitness value starts increasing. A reverse-interval PLS can be implemented where the first model is developed with all intervals and intervals are removed one at a time until the lowest fitness value is reached. Note that the number of PCs to use in each model is chosen by the user at the beginning of the optimization process. The main disadvantage of interval PLS is its calculation time. When the size of intervals is low and the number of variables large, the convergence time can get very long (hours).

7.1.1.2. Genetic algorithms

Alternatives to exhaustive search methods have been developed based on stochastic approaches. A **stochastic method** is a method involving some randomness. Genetic algorithms are stochastic methods that imitate biological evolution ("the best individual survives"). They involve gene modification methods (combination and mutation) to progressively improve the breed, in this case the fit (i.e. decrease the predictive error). To use genetic algorithms for variable selection, it is necessary to randomly determine individuals: a subset of consecutive (or not) variables. PLS or PCR models are developed using each individual in calibration and RMSECV is used to evaluate the fit. The number of factors to use is predetermined by the user. At the end of this first step, a table representing, for each individual, its fitness and the variable ID is obtained. The flag yes/no of each variable that determine if a specific variable is part of an individual is called a gene. Table 2 presents such map of individual/gene.

The second step of the algorithm is discarding individuals that have a fitness value (RMSECV) larger than the median. Noisy variables (responsible for higher RMSECV) are discarded through the whole process, because less and less individuals have them. In the present example, individuals 1, 3, 4, 5, and 10 will be discarded. To "repopulate" the number of individuals, there are combinations or cross-over among the survival individuals. The variability in the population is generated by mutations (the frequency of mutation is usually user-defined).

Table 2. Map of individuals with selected variables of a 30 wavelength spectrum (white squares represent selected variable).



For practicality, we will represent black squares with 0 and white squares with 1. Individuals 1 and 2 from the original population can then be represented as follows:

Individual 1: 01011011101110011101110110011

Individual 2: 11011110111101110111011101111

A combination or crossover consists in exchanging parts of the parent genes to create new individuals. The crossover can be single or double. In the single crossover, a variable is selected and everything on the left of this variable will be crossedover to the other individual. In a double crossover, two parts of the individual are exchanged. Thus, using Individuals 1 and 2 as parents, we would obtain with a single cross-over at variable 21, the following two new individuals:

New Individual 11: 010110111011100111001|011101111

New Individual 12: 110111101111011101111|110110011

and with a double cross-over at variables 7 and 15, we would obtain:

New Individual 11: 0101101|01111011|111001110110011

New Individual 12: 1101111|11011100|101111011101111

If only combinations were possible, the offspring could not contain variables that were not selected during the initial random assignment. Mutations allow this by randomly converting a non-selected variable to a selected variable. Using individual 1 as parent, a mutation occurring at gene 10 would give:

New Individual 11: 0101101111**1**11001110011110110011

The generation of offspring carries on until the population of individuals is back to its original size (10 in our example). RMSECVs are calculated, just like after the random initial population assignment. The optimization process carries on as long as some percentages of individuals use the same variables, the total number of generation is reached (also set by the user), or the optimization goal is achieved. At the end, the

selected variables can then be used to develop a calibration model, where the number of principal components can be chosen. The cross-validation process can also be replaced by an independent validation set to evaluate the real performance of the model. The whole process is thus summarized in 5 steps: 1. random creation of individuals, 2. fitness evaluation, 3. individual removal, 4. population breeding, 5. population mutation. Steps 2 to 5 are carried on until the algorithm stops.

The use of genetic algorithms requires the tuning of many parameters (the number of PCs to use in the automatically generated models, the nature of combination –there are different approaches- , the rate of mutation, the number of generations allowed …) and it is also prone to overfitting. A consistent validation strategy should be implemented to avoid that. Genetic algorithms have been used for twenty years in NIRS, but the main issue that developers are facing is that few instrument have software able to implement them (only available in statistical suites). However, they can be easily applied into instruments because regression coefficients of non selected variables can simply be set to zero. The literature reports successful applications over full spectrum methods for soluble solid content in apple (Shi et al, 2008) and grains (Davies, 1987; Leardi et al, 1992).

7.1.1.3. Particle swarm optimization

Another stochastic approach for variable selection is particle swarm optimization (PSO). PSO represents a family of algorithms that mimic the behavior of social insects or swarms. They were introduced by James Kennedy and Russell Eberhart in 1995 and are based on three sociocognitive underpinnings: Evaluate, Compare, and Imitate. For variable selection, the use of PSO is quite similar to GAs since it is a binary combinatory problem. PSO has been used for the optimization of ANN (to replace the back propagation algorithm) and this technique generates a lot of interest in the NIRS community (Kennedy and Eberhart, 2001).

In PSO, each particle (individual in GA) – a subset of variables among the variable population – is provided with an initial velocity. At each generation, the fitness of each particle is compared and its velocity is updated in function of its performance at generation n and earlier, the best global particle, and the best neighboring particles. Thus,

the best performing particles will be imitated by other particles and the process will carry on until the performance criterion is reached. The pseudo code of PSO is provided below.

```
1-  Loop
2-      For i = 1 to Number of Particle
3-          If Fitness(Particle i)  < Fitness(Historical Best)
4-              For d = 1 to Dimensions
5-                  Particle_id = New Best_id
6-              Next d
7-          End If
8-          g = i
9-          For j = indexes of Neighbors
10-             If Fitness (Neighbors j) < Fitness (New Best)
11-                 Then g = j
12-             End If
13-         Next j
14-         For d = 1 to number of Dimensions
15-             v_id(t) = v_id(t-1) x W_i + φ1 x rand() x (New Best_id – …
16-               Particle_id (t-1)) + φ2 x rand() X (Best Neighbor_gd – …
17-               Particle_id (t-1))
18-             If ρ_id < s(v_id(t))
19-                 Then Particle_id (t) = 1
20-                 Else Particle_id (t) = 0
21-             End If
22-         Next d
23-     Nest i
24- Until criterion
```

Step 1: Adapt based on best overall

Step 2: Find best in neighbors

Step 3: Update velocity and particles

where v_{id} represents the velocity of the particle i and dimension d at step t or t-1, W_i is the inertia weight for the particle *i*, φ1 and φ2 are learning factors whose sum is 2.0, and ρ is

a random vector between 0 and 1. $v_{id}(t)$ is the parameter that determines if a dimension *d* will take the value 0 or 1 (0 meaning the variable is not selected, and 1, the variable is selected).

To obtain values between 0 and 1 the velocity is scaled using a sigmoid function.

Equation 18. $$s(v_{id}) = \frac{1}{1+\exp(-v_{id})}$$

Since PSO looks for the best performing particles in the entire population and best previous state of the particle, the convergence is often quicker than for GA. The fitness parameter is usually the RMSECV. Since few studies have used PSO for variable selection, it is not possible to tell which method, PSO or GA, performed better for NIR spectra. More theory on PSO can be found in Kennedy and Eberhart (2001).

7.1.1.4.Other methods

Other optimization methods have been implemented to select variables. We can mention simulated annealing and evolutionary programming. We will not discuss these techniques, but their use is increasing in NIRS (Luke, 1994; Swierenga et al, 1998; Lu et al, 2004; Shen et al, 2004).

7.1.2. Local regression

Local models aim to optimize the prediction model statistics by choosing only the most appropriate samples to include in the calibration set. A model is generated each time a new sample is scanned for prediction. Tom Fearn (2001) stated that “A local calibration is one in which the equation used to predict for a given unknown sample is derived from only a subset of the available training samples, this subset having been chosen because the spectral data for the samples it contains resembles the spectral data for the unknown in some particular way”.

Local models have been developed to deal with clustering issues or non linearity issues. Figure 38.A presents the relationship between a parameter, y, and a univariate **X**. A clear non-linear pattern exists. However, as displayed by figure 38.B, it is possible to select only a reduced range of data, and still be able to use linear regression methods to perform an accurate prediction over this range.

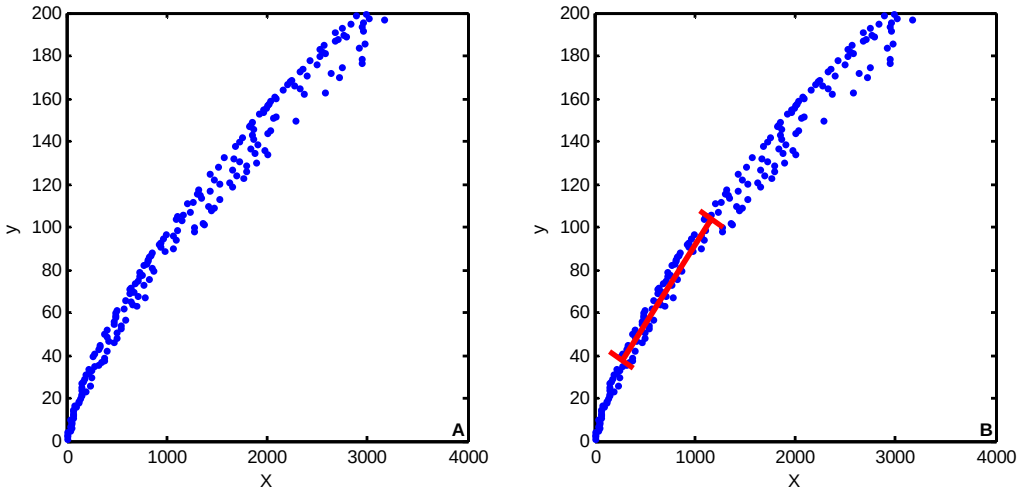


Figure 38. Non-linearity issues. Plot **A** shows a non-linear relationship between two variables X and y. Plot **B** presents a possible data range where local regression with linear regression method is possible.

The implementation of local models globally follows the same steps, no matter the algorithm used:

1. Creation of a calibration set
2. Collection of a new sample spectrum to be predicted
3. Selection of samples the “closest” or the most “similar” from the calibration set to the sample to predict
4. Prediction of the new sample based on the “local” model.

First, it is necessary to have a large library of samples. Enough samples should be present in the neighborhood of the new samples to ensure its precise and accurate prediction. It is then necessary to select the most appropriate samples from the available database to develop a local calibration set. Depending on the algorithm, different methods are used to find the closet samples. Here, the three most common techniques used in NIRS will be presented (CARNAC, LOCAL, and locally weighted regression).

In the method CARNAC (Davies et al, 1988; Davies and Fearn, 2006b), standing for Comparison Analysis using Restructured Near-infrared And Constituent data, the database, as well as the new sample to predict, is compressed using Fourier or wavelet transform and a similarity index is determined. This index is defined as

Equation 19. $$s = \frac{1}{(1 - r^2)}$$

where $\mathbf{r^2}$, the coefficient of determination, exposes the similarity of the new sample with those in the database. The number of samples to include can be set by either using a threshold on **s** or by selecting a fixed number of samples, among the closest. The algorithm LOCAL (Shenk et al, 1997) performs similarly by determining the correlation coefficient (r) between the spectra in the database and the new spectrum (US Patent 5798526 – August 1998). Spectra can be raw or pretreated, but no compression method is used. Similarly, the number of samples to keep is determined by a threshold or a fixed number of samples to include.

The locally weighted regression (LWR) algorithm (Cleveland et al, 1988, Næs et al, 1990) uses a different approach to find the closest spectra. A PCA compresses the database and the sample to predict. The closest samples are determined by calculating Euclidian or Mahalanobis distances between scores of the new sample and scores of the calibration set. A modification to LWR (LWRY) introduced the possibility to add information from the Y-matrix in the distance calculation (Chang et al., 2001).

The Mahalanobis distance was replaced with a more complex distance expression:

Equation 20. $$D_i = \alpha Yd_i + (1 - \alpha) Xd_i$$

where Xd_i is the Mahalanobis distance for sample i to predict, α is a weighting factor and Yd is a normalized chemical distance calculated as follows:

Equation 21. $$Yd_i = \frac{|y - y_{i,l}|}{\max(|y - y_{i,l}|)}$$

where y, the estimated unknown property estimated by a global regression and $y_{i,l}$ is the reference value of each local sample.

The next step consists of developing the local model and predicting the new sample. CARNAC is particular because it does not use any regression method, but the prediction of the new sample is determined by a weighted average of the y values of the selected samples. The logarithm of the similarity index is used as a weighting factor. LOCAL and LWR methods, however, predict the y value of the new spectrum using a regression method (PLS or PCR). In LOCAL, several models are developed for different PCs and the final prediction is a weighted average of all predictions. LOCAL algorithm automates the choice of the number of PCs. The weight associated with each prediction is:

Equation 22. $$W_k = \frac{1}{X_{residual} \times RMS_{BetaCoef}}$$

where W_k is the weight associated with the kth PC, $X_{residual}$, the difference between the reconstructed spectra (from PLS or PCR decomposition) and the true spectrum to be predicted, and $RMS_{BetaCoef}$, the root mean square of the regression coefficient. The best principal component will present the best reconstructed spectra and the smoothest

regression coefficient and will thus maximize the weighting factor ($X_{residual}$ decreases while $RMS_{BetaCoef}$ increases with the number of PCs). For LWR, the determination of the best PC is not automated. However, a weighting function is applied to each selected sample using a cubic weight function (figure 39) based on the distance of the sample to predict with the local calibration samples (the farthest is scaled to have a distance of 1). The weighting function is as follows:

Equation 23. $$W_S = \left(1 - d^3\right)^3$$

where W_S is the weight associated with a calibration sample and *d*, the scaled distance between the new sample to predict and the calibration sample. Like every calibration procedure, the last step in the development of a local method should be a validation procedure.

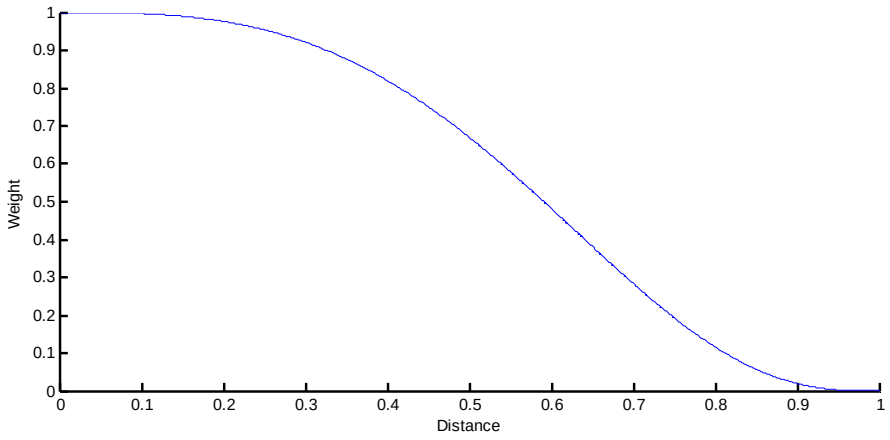


Figure 39. Cubic weight function.

Local strategies have their pros and cons. They can be as accurate as global models, robust, and able to deal with non-linear relationships. However, they are time consuming, have little instrumental support (mostly offline applications), and cannot be approved in heavy regulatory environments. From a local model point of view, no

calibration can be approved because each model is different. However, this regulatory issue could be solved by approving the sample database, since the database is fixed for each new sample (Cogdill et al, 2002).

7.2) Standarization

Individual calibrations can be developed for each instrument, but sometimes the resources are limited. Then, calibrations developed from a single instrument can be transferred to the rest. Calibration models are often developed using spectra of a single instrument called a master unit. This master unit should be chosen among other units of the network of instrument available to the chemometrician because of its precision and accuracy (Siska et al, 2001). However, the master unit is often located in the research laboratory and does not participate to the actual online processes or quality controls. The model needs to be transferred to these instruments. This is called standardization and is an essential part of the implementation of NIRS.

Three types of standardization techniques exist:

- Optical standardization or calibration transfer by adaptation of the secondary unit's spectra to match master's spectra
- Post regression correction or standardization by correction of the predicted values of the secondary units using calibration developed on master unit.
- Robust models or standardization by calibration model adaptation

7.2.1) Optical standardization techniques

Optical standardization techniques try to adjust secondary unit's spectra to the spectra of the master unit. The intent is for the calibration model to perform as well as if a sample was scanned on the master unit. To perform the spectral modifications, a set of samples, called standardization samples, need to be run on both master and secondary units. Bouveresse and Massart (1996) presented a comprehensive description of the

selection and the use of standardization samples in their review of standardization techniques. This can be done with a high leverage method proposed by Wang et al (1991) or the Kennard-Stone algorithm (Kennard et al, 1969). Bouveresse et al (1994) presented two simple techniques to select standardization samples. The first consisted in selecting samples that were the best predicted among the validation set. The second consisted in using samples from a different source but of similar nature (not artificial samples). When dealing with several prediction models, using the same standardization samples for all factors is often convenient, but may not result in the best standardization since a sample may be predicted well for a factor but not for another.

Several techniques exist to match spectra. Five of the best known techniques will be presented here. Other methods have been developed, but remain marginal in their use because their lengthy calculations do not bring real improvements compared to existing methods. Examples of complex techniques are ANN based standardization (Despagne et al, 1998; Duponchel et al, 1999), maximum likelihood PCA (Andrews et al, 1997), positive matrix factorization (Xie et al, 1999), Kalman and Wiener filters (Teppola et al, 1999 and Siska et al, 2001 respectively), and other standardization of the regression coefficient techniques (Wang et al, 1991).

7.2.1.1 Single wavelength standardization

This method, developed by Shenk and Westerhaus (Paynter et al, 1983; Shenk et al, 1985), consists of correcting, for each wavelength, the absorbance shifts (adjust the intensity differences between instruments). It is done by regressing, one at a time, the absorption values of a standardization set scanned on the master unit against the one obtained when the same set is scanned on the secondary unit. A slope and an offset are obtained and the secondary unit is corrected. This simplistic technique is applicable only when a limited shift in the **X**-axis (wavelength) exists.

Figure 40 shows the application of this technique in standardizing two Foss Infratec units (master unit: Infratec 1241, secondary unit: Infratec 1229). Twenty soybean standardization samples were used and the wavelength of interest was 852 nm. The

secondary unit had a lower intensity in its measurements and the correction function at this wavelength would be:

Equation 24. $$W_{Scor}^{852} = 0.88 \times W_{S}^{852} + 0.19$$

where W_{Scor}^{852} corresponds to the corrected value of the secondary unit at 852 nm, W_{S}^{852} is the original value of the secondary unit, and 0.88 and 0.19 are the slope and the offset values respectively obtained when fitting a first order polynomial to the data.

The quality of the standardization samples is critical because a wrong approximation of the correction factors at each wavelength can have a large impact on the final SEP.

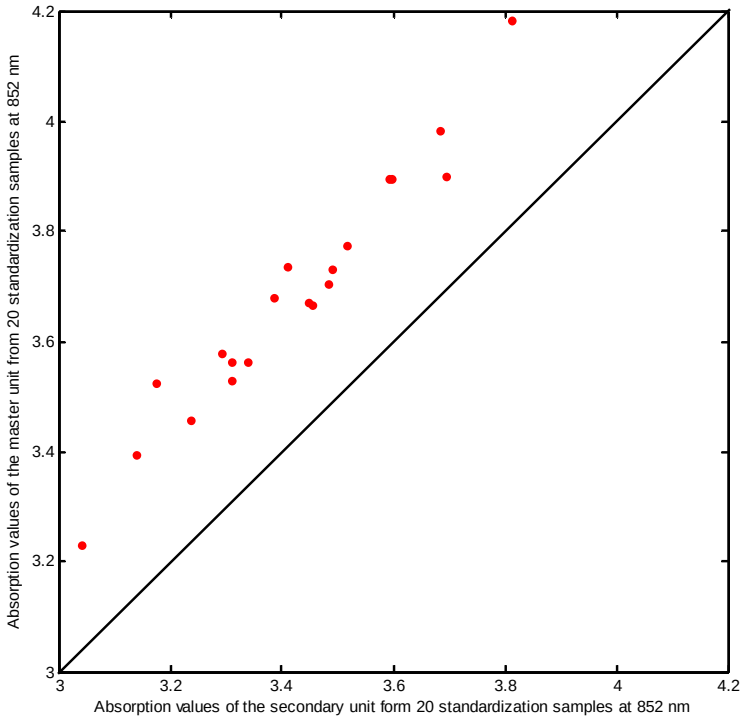


Figure 40. Single wavelength standardization. An offset and a slope need to be applied to this wavelength to correct the signal.

7.2.1.2. The patented algorithm – Shenk and Westerhaus

An updated version of the single wavelength standardization included the possibility to correct for shifts in the **X**-axis. This method has been patented and is used by the chemometric package WinISI (US Patent 4866644 – September 1989). Shenk et al (1993) and Bouveresse et al (1994) describe the functioning of the algorithm.

It is divided into two steps. The first step is called wavelength index correction and consists in correcting the **X**-axis shifts while the second part corrects the intensity differences. Standardization spectra are preprocessed using a first derivative treatment. For each wavelength of the master, a spectral window of neighboring wavelength on the secondary instrument is chosen and for each wavelength of the window, the correlations with the master are computed. A quadratic model is fit to the correlation values to estimate more precisely the position of the wavelength that produces maximum correlation. The fit is between the wavelength that has the highest correlation and its two neighboring wavelengths. The new locations obtained from the quadratic model (inflection points of the windowed quadratic equation) are recorded and a new spectrum is built. A second quadratic model is developed to relate the master wavelength to the matching wavelength on the modified secondary unit spectra. Definitive values for the secondary unit wavelengths corresponding to the master wavelengths are obtained. This process is called the wavelength index correction. The modified secondary unit spectra are then interpolated. Wavelengths of the secondary unit are shifted to the corresponding master wavelength, at each wavelength (similar to what was implemented in the single wavelength standardization method) using a linear regression. A slope and an offset for each wavelength are obtained and are used to correct spectral intensity.

The wavelength index and the spectral intensity correction factors are stored and applied to new spectra scanned on the slave instrument. This method is adapted to many standardization situations involving similar instruments. When facing instruments with more complex differences like a peak broadening, the patented algorithm is not sufficient because it “assume that no relationship exists between neighboring correction models” (Feudale et al, 2002).

7.2.1.3. Direct standardization

Wang et al (1991) introduced two standardization techniques able to cope with peak broadening. Both methods imply that spectra from the master and the secondary unit are linearly related and this relation can be described by a transformation matrix such as

Equation 25. $\mathbf{S_M = S_S F}$

where $\mathbf{S_M}$ is a spectra scanned on the master unit, $\mathbf{S_S}$ a spectra scanned on the secondary unit, and $\mathbf{F}$ the transformation matrix. In direct standardization (DS), the transformation matrix is simply estimated as

Equation 26. $\mathbf{F = S_S^{+} S_M}$

where $\mathbf{S_S^{+}}$ is the pseudo inverse of $\mathbf{S_S}$. With $\mathbf{F}$ calculated, any new spectra $\mathbf{S_N}$ can be modified to the original measurement space so that the calibration will predict it appropriately using

Equation 27. $\mathbf{S_{Ntrans} = S_N F}$

where $\mathbf{S_{Ntrans}}$ is the new spectrum modified to the original measurement space. $\mathbf{S_{Ntrans}}$ is then applied to the calibration model developed on the master unit.

The computation of $\mathbf{F}$ assumes that the difference between instruments is due to instrumental variations. However, the variation in the chemical composition (due to non-complete sample homogeneity) is also modeled and may be a source of error. Also, because the number of samples used to create $\mathbf{F}$ (standardization samples) is smaller than the number of channels to evaluate, DS is subject to overfitting. $\mathbf{F}$ is typically estimated using PCR or PLS to obtain a least squares solution.

Another approach to limit overfitting is to reduce the number of channels estimated at a time. This is the purpose of piecewise direct standardization (PDS).

7.2.1.4.Piecewise direct standardization

In DS, each wavelength of the master unit is related to all the wavelengths of the secondary unit. PDS is a local alternative to DS. Influence of shifts of the **X**-axis between instruments is limited to certain spectral regions and does not impact the standardization of other wavelengths. PDS performs the same calculations as DS but at a local level. The user defines a window size and transformation coefficients are calculated to relate one wavelength of the master unit with several wavelengths on the secondary unit:

Equation 28. $$\mathbf{r}_j = \mathbf{R}_j \mathbf{b}_j$$

where $\mathbf{r}_j$ is the absorption value at wavelength j from the master unit, $\mathbf{R}_j$ is the absorption value at wavelength j from the secondary unit, and $\mathbf{b}_j$ is the vector of transformation coefficients for wavelength j. The size of the window usually varies between 3 and 5 wavelengths and a wavelength on the secondary unit is used several times to explain different wavelength on the master unit (moving window, one master wavelength at a time). For each window, transformation coefficients are estimated and assembled to form a banded diagonal matrix **F** according to

Equation 29. $$\mathbf{F} = diag\left(\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2^{\mathrm{T}}, ..., \mathbf{b}_j^{\mathrm{T}}, ..., \mathbf{b}_k^{\mathrm{T}}\right)$$

where k is the number of wavelengths. Thus, the transfer matrix relates the response of a number of wavelengths (size of the window) of the secondary unit to a single wavelength of the master unit (figure 41).

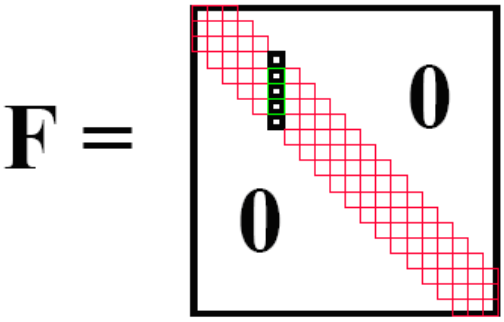


Figure 41. Structure of the transformation matrix

PDS is often used as reference method for other standardization techniques. Many authors have used it successfully, but few instruments are able to use a transfer matrix onboard (Perten DA (Perten Instruments AB, Huddinge, Sweden) can use PDS with a window size of 1). The standardization using DS or PDS has to be done offline. This situation is similar to the use of the patented algorithm except that Foss NIRSystems instruments are equipped with software that can accommodate the standardization parameters. Notice that a PDS with a window size of 1 is equivalent to single wavelength standardization. The difference is that in PDS, coefficients are estimated using a PCR or PLS.

To carry on the example introduced with single wavelength standardization, table 3 presents the transformation matrix for wavelength at 850 nm, 852 nm, and 854 nm when a window size of 3 is used on the secondary unit. With each wavelength a bias is associated. In this particular case, they are -0.0114, -0.0074, and -0.0077 for 850 nm, 852 nm, and 854 nm respectively. With these parameters, the calculation of the corrected absorption value of a spectrum collected on a secondary unit for the wavelength at 852 nm would be:

Equation 30. $$W_{Scor}^{852} = 0.3601 \times W_{Sraw}^{850} + 0.3601 \times W_{Sraw}^{852} + 0.3589 \times W_{Sraw}^{854} + (-0.0074)$$

The reason for coefficients to be so similar is that there is no shift in the wavelength axis.

Table 3. Structure of the transformation matrix for the example using a window size of 3 wavelengths.

		Master Unit			
		850	**852**	**854**	**…**
Secondary Unit	**850**	0.5420	0.3601		
	852	0.5401	0.3601	0.3614	
	854		0.3589	0.3601	…
	856			0.3589	…
	…				…

An attempt to use DS and PDS on spectra transformed in the wavelet domain was published by Tan et al (2001). They showed that the standardization at the approximation level with PDS and the detail level with DS was more robust and reliable than using DS or PDS only.

7.2.2) Post regression correction techniques

This second type of model transfer method does not modify the spectra. It adjusts predictions made on the secondary unit by the model developed on the master unit. Post regression correction techniques are easily implemented by every embedded software or firmware and are also easier to understand (for users).

Often called slope and bias correction, this technique is widely used among very similar instruments (one can note the misuse of the term bias which in reality corresponds to term intercept). In this technique, a standardization set is scanned on all instruments (the secondary units as well as the master units) and predictions are compared to the reference values (Bouveresse et al, 1996). A linear regression is implemented and the

slope and intercept of the trend line are determined. These parameters are then used to systematically correct predictions from the secondary instrument using the prediction equation developed on a master unit. This is a very simple approach that requires only that the range of the standardization samples cover the range of the calibration samples.

A special case of slope and intercept correction is bias only correction. Bias is calculated from standardization samples predictions and reference values and is added to the future predictions. This situation is suitable only when the slope is significantly equal to one. This signifies that the difference between the two units is mainly based on absorption intensities differences and not a shift in the **X**-axis. The slope in post regression correction appears when the wavelength alignment between the instruments is not correct. As an example, predictions of the standardization set by a protein model developed on Infratec 1241 and run on Infratec 1229 are compared with the reference measurements. Figure 42 shows the relationship between the two units as well as the slope and intercept necessary to implement slope and bias correction and the bias for bias only correction.

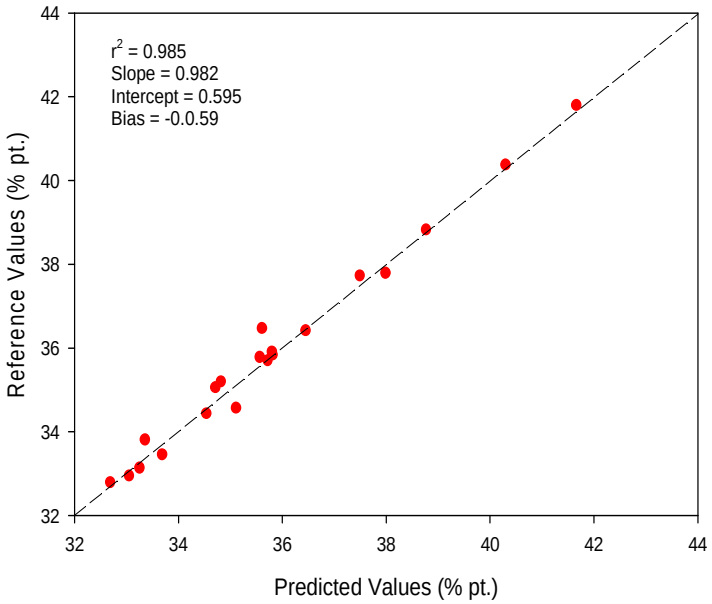


Figure 42. Post-regression correction results for protein. Slope and intercept or bias only correction can be used to correct the measurements.

In this situation, using slope and bias post regression correction, the prediction on the secondary unit should be corrected as follows:

Equation 31. $$\hat{X}_{Cor}^{Sec} = 0.982 \times \hat{X}_{Raw}^{Sec} + 0.595$$

And using bias only correction, it would look like:

Equation 32. $$\hat{X}_{Cor}^{Sec} = \hat{X}_{Raw}^{Sec} + (-0.059)$$

7.2.3) Robust standardization techniques

Robust statistics are statistics that do not rely as much as classical statistics on assumptions. Many of the classical techniques require a normal or uniform distribution of the data. It is the case of MLR, PCR, and PLS while ANN and SVM are robust methods. They are more tolerant to outliers and emulate classical methods. However, the term robust, in robust standardization techniques, represents the idea of being resistant to outliers, but does not always involve the use of robust statistic techniques. Several robust standardization methods have been developed, but few papers have been published on the topic.

7.2.3.1. Spectral preprocessing

A possibility to make a calibration robust is to use pretreatment methods to remove instrumental interferences. But these methods do not work if there is a shift in the **X**-axis among instruments (Perten DA instruments uses MSC to standardize diode array instruments). Standardization by preprocessing is fairly easy, but cannot cope with large or inconsistent instrumental differences.

7.2.3.2. Variable selection

It is also possible to make a model robust by selecting wavelengths that are not subject to shifts (isonumeric wavelengths (Mark and Workman, 1988)) and to develop models using MLR. However, it is necessary to correct for intensity differences (Dean and Isaksson, 1993).

7.2.3.3. Including external variability

This family of robust methods consists in scanning all or part of the calibration set on all the instruments of the network and by developing a calibration based on all the spectra. This way, the calibration algorithm does the standardization work by adapting the model to each instrument. Fearn (2001) noted that "The resulting calibration will almost certainly be less accurate than a calibration for a single instrument, but this may be a price worth paying". Often preprocessing methods are used to help remove the instrumental interferences.

7.2.3.4. Orthogonal-based techniques

These methods are based on the statistical theory that the column space of the **X**-matrix (set of all possible linear combination of column vectors) is "the sum of two subspaces, among which only one contains information useful for the model "(Roger et al, 2003). Thus, by doing the appropriate projection, one can develop a model based only on the adequate **X**-matrix. There exists two ways of estimating the "parasitic" subspace. The first consists of finding the space orthogonal to y. This is what orthogonal signal projection (OSC) and orthogonal projections to latent structures (O-PLS) are doing. The second approach estimates the space in which the external factor occurs. Transfer by orthogonal projection (TOP), dynamic orthogonal projection (DOP), and error removal by orthogonal subtraction (EROS) are the most popular techniques to find and remove the orthogonal space.

Orthogonal signal correction. OSC was introduced by Wold et al in 1998. The idea is to compute loading weights, **w**, such that the scores **t** calculated from **t** = **Xw**

describe as much variance as possible. The decomposition into scores is similar to the one used to develop PLS. The constraint is that this variance is not correlated to y. It is possible to repeat the process and remove more orthogonal factors from the **t** matrix. This corrected **t** matrix is then used to develop a PLS model.

Orthogonal projections to latent structures. Trygg and Wold introduced a variation of OSC in 2002. Instead of removing the orthogonal signal from the **t** matrix prior to calibration development, the O-PLS method removes the orthogonal information from the PCs calculated by PLS. The reconstructed **X**-matrix can then be used to develop a calibration model. O-PLS was reported to perform better than OSC. Their primary goal is to remove noise. Authors reported that systematic variations could be successfully removed by the methods. These variations can be due to scattering or baseline effects, but from a standardization standpoint, the removal of systematic variations can correspond to deleting instrumental signatures and thus improve model transferability.

Transfer by orthogonal projection. Andrew and Fearn (2004) published an orthogonal technique aiming to remove the interfering components from the **X**-matrix in the situation of calibration transfer. Standardization samples are measured on both master and secondary units. A difference spectrum is obtained by subtracting standardization samples from both units and a PCA is performed on the difference matrix **D**. The first *k* loadings of **D** are used to form the matrix **P** representing the direction of main variation between units. **P** is orthogonalized on the **X**-matrix to obtain a corrected **X**-matrix ($\mathbf{X_{corr}}$) for between-instrument differences using

Equation 33. $$X_{corr} = X\left(I - PP^{T}\right)$$

where **I** is the identity matrix. $\mathbf{X_{corr}}$ is then used to develop calibration models. TOP requires the use of standardization samples. But in certain situations, collection of samples is not possible and orthogonalization is performed using the DOP method.

Dynamic orthogonal projection. In DOP (Zeaiter et al, 2006), standardization samples are replaced by virtual standards. These standards are chosen to represent a variation to be removed. In calibration transfer situations, a few samples collected from the secondary instruments (X_t) are used to create virtual standards using a kernel function. Equation 51 demonstrates how virtual standards are created using the calibration set (**X** and y) and the reference value of the samples collected on the secondary unit (y_t):

Equation 34. $$\hat{X}_t = AX \text{ with } a_{ij} = F_{y_{ti}}\left(y_j\right)$$

where $\hat{X}_t$ are the virtual standards, $F_{y_{ti}}$ is a Gaussian kernel function centered on y_{ti} for the ith sample and the jth variable. A difference matrix **D** is calculated ($D = \hat{X}_t - X_t$) and similarly to TOP, the first *k* loadings from a PCA are selected to form **P**. The **X**-matrix is then orthogonalized similarly to TOP using equation 34. By not requiring the same samples run on both master and secondary unit, DOP is more flexible, especially for on-line applications.

Error removal by orthogonal subtraction. EROS (Zhu et al, 2008) is based on the same principles as TOP and DOP. The difference arises from the way the **P** matrix is calculated. In EROS, the difference matrix **D** is mean centered and a new matrix **W** representing the difference between measurements is calculated as follows:

Equation 35. $$W = \sum_{i=1}^{m} \frac{D_i D_i^T}{(r-m)}$$

where **r** is the total number of spectra in the **D** matrixes and *m* is the number of samples (also the number of **D** matrixes). **W** represents the pooled within-sample covariance matrix of the replicate spectra. The *k* first PCA loadings of **W** are selected and used to form the matrix **P**, used to orthogonalize the **X**-matrix using equation 34.

Orthogonal methods need to be tuned. The number of times orthogonal components are removed for OSC and O-PLS need to be optimized. Similarly, TOP, DOP, and EROS require tuning *k*, the number of factors to orthogonalize.

TOP, DOP, and EROS have the advantage of being embedded in the calibration model (the regression algorithm will not take into account the orthogonal signal and the beta coefficients will emphasis the orthogonalization by not taking into account the orthogonal part of the signal in prediction mode). Thus, these methods are implemented during the calibration process. Validation samples do not need any orthogonalization step on the contrary to all other preprocessing and standardization methods. They have the tremendous advantage to be useable by all instrument software settings (even firmware). These same methods can be used to remove other interferences than instrumental differences such as temperature, batch effects, and other unwanted external effects presenting a repetitive effect across samples (Zeaiter et al, 2006).

8. THE UNSCRAMBLER EXAMPLES

8.1) Carrying out Principal Component Analysis (PCA)

8.1.1) Importing data

For this example you will need the data file “Pipes.csv”. This file cannot be open directly in The Unscrambler, it is a comma delimited file. You need to import it. So go to ***file → import → ascii*** and then browse for the file. You can open the file in Excel so then you can see how it looks like, and fill the right fields from the importing window (figure 43).

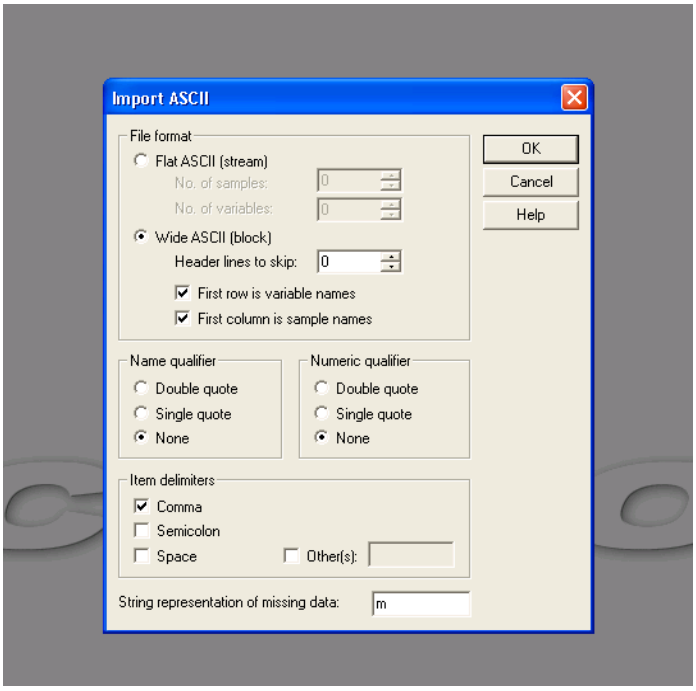
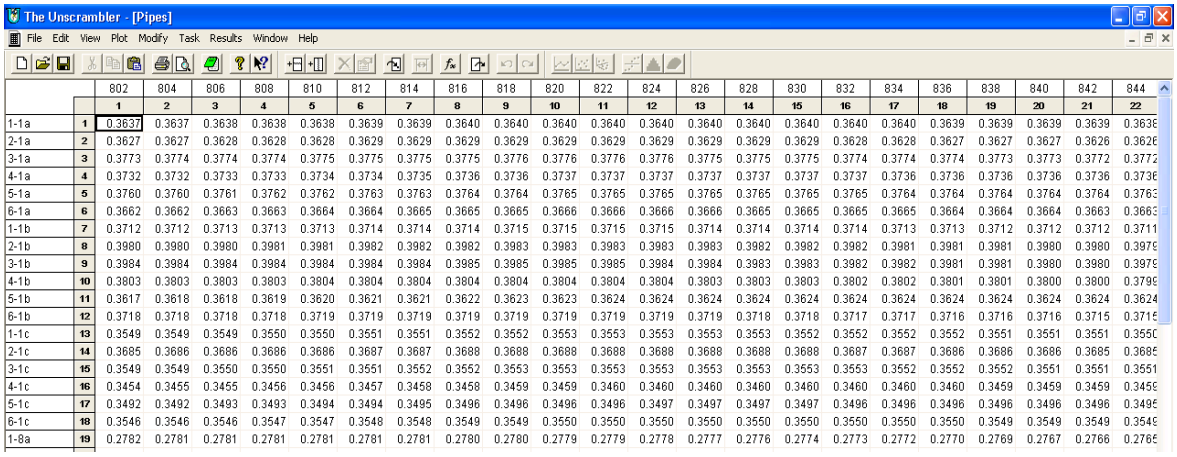


Figure 43. Importing ascii window

Once the data is imported, the data table should have the wavelength number on the top row and the sample numbers on the first column (Figure 44). The first wavelength (top row) is 802 nanometers and the last one is 2498 nm, so you can suspect that this is data from a reflectance instrument. The instrument reads the absorbance every 2 nm (check how wavelengths increase by two nanometers).



		802	804	806	808	810	812	814	816	818	820	822	824	826	828	830	832	834	836	838	840	842	844
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1-1a	1	0.3637	0.3637	0.3638	0.3638	0.3638	0.3639	0.3639	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3640	0.3639	0.3639	0.3639	0.3639	0.3638
2-1a	2	0.3627	0.3627	0.3628	0.3628	0.3628	0.3629	0.3629	0.3629	0.3629	0.3629	0.3629	0.3629	0.3629	0.3629	0.3629	0.3628	0.3628	0.3627	0.3627	0.3627	0.3626	0.3626
3-1a	3	0.3773	0.3774	0.3774	0.3774	0.3775	0.3775	0.3775	0.3775	0.3776	0.3776	0.3776	0.3776	0.3775	0.3775	0.3775	0.3774	0.3774	0.3774	0.3773	0.3773	0.3772	0.3772
4-1a	4	0.3732	0.3732	0.3733	0.3733	0.3734	0.3734	0.3735	0.3736	0.3736	0.3737	0.3737	0.3737	0.3737	0.3737	0.3737	0.3737	0.3736	0.3736	0.3736	0.3736	0.3736	0.3736
5-1a	5	0.3760	0.3760	0.3761	0.3762	0.3762	0.3763	0.3763	0.3764	0.3764	0.3765	0.3765	0.3765	0.3765	0.3765	0.3765	0.3765	0.3764	0.3764	0.3764	0.3764	0.3764	0.3763
6-1a	6	0.3662	0.3662	0.3663	0.3663	0.3664	0.3664	0.3665	0.3665	0.3665	0.3666	0.3666	0.3666	0.3666	0.3665	0.3665	0.3665	0.3665	0.3664	0.3664	0.3664	0.3663	0.3663
1-1b	7	0.3712	0.3712	0.3713	0.3713	0.3713	0.3714	0.3714	0.3714	0.3715	0.3715	0.3715	0.3715	0.3714	0.3714	0.3714	0.3714	0.3713	0.3713	0.3712	0.3712	0.3712	0.3711
2-1b	8	0.3980	0.3980	0.3980	0.3981	0.3981	0.3982	0.3982	0.3982	0.3983	0.3983	0.3983	0.3983	0.3983	0.3982	0.3982	0.3982	0.3981	0.3981	0.3981	0.3980	0.3980	0.3979
3-1b	9	0.3984	0.3984	0.3984	0.3984	0.3984	0.3984	0.3984	0.3985	0.3985	0.3985	0.3985	0.3984	0.3984	0.3983	0.3983	0.3982	0.3982	0.3981	0.3981	0.3980	0.3980	0.3979
4-1b	10	0.3803	0.3803	0.3803	0.3803	0.3804	0.3804	0.3804	0.3804	0.3804	0.3804	0.3804	0.3804	0.3803	0.3803	0.3803	0.3802	0.3802	0.3801	0.3801	0.3800	0.3800	0.3799
5-1b	11	0.3617	0.3618	0.3618	0.3619	0.3620	0.3621	0.3621	0.3622	0.3623	0.3623	0.3624	0.3624	0.3624	0.3624	0.3624	0.3624	0.3624	0.3624	0.3624	0.3624	0.3624	0.3624
6-1b	12	0.3718	0.3718	0.3718	0.3718	0.3719	0.3719	0.3719	0.3719	0.3719	0.3719	0.3719	0.3719	0.3719	0.3718	0.3718	0.3717	0.3717	0.3716	0.3716	0.3716	0.3715	0.3715
1-1c	13	0.3549	0.3549	0.3549	0.3550	0.3550	0.3551	0.3551	0.3552	0.3552	0.3553	0.3553	0.3553	0.3553	0.3553	0.3553	0.3552	0.3552	0.3552	0.3551	0.3551	0.3551	0.3550
2-1c	14	0.3685	0.3686	0.3686	0.3686	0.3686	0.3687	0.3687	0.3688	0.3688	0.3688	0.3688	0.3688	0.3688	0.3688	0.3688	0.3687	0.3687	0.3686	0.3686	0.3686	0.3685	0.3685
3-1c	15	0.3549	0.3549	0.3550	0.3550	0.3551	0.3551	0.3552	0.3552	0.3553	0.3553	0.3553	0.3553	0.3553	0.3553	0.3553	0.3553	0.3552	0.3552	0.3552	0.3551	0.3551	0.3551
4-1c	16	0.3454	0.3455	0.3455	0.3456	0.3456	0.3457	0.3458	0.3458	0.3459	0.3459	0.3460	0.3460	0.3460	0.3460	0.3460	0.3460	0.3460	0.3460	0.3459	0.3459	0.3459	0.3459
5-1c	17	0.3492	0.3492	0.3493	0.3493	0.3494	0.3494	0.3495	0.3496	0.3496	0.3496	0.3496	0.3497	0.3497	0.3497	0.3497	0.3496	0.3496	0.3496	0.3496	0.3496	0.3496	0.3495
6-1c	18	0.3546	0.3546	0.3546	0.3547	0.3547	0.3548	0.3548	0.3549	0.3549	0.3550	0.3550	0.3550	0.3550	0.3550	0.3550	0.3550	0.3550	0.3550	0.3549	0.3549	0.3549	0.3549
1-8a	19	0.2782	0.2781	0.2781	0.2781	0.2781	0.2781	0.2781	0.2780	0.2780	0.2779	0.2779	0.2778	0.2777	0.2776	0.2774	0.2773	0.2772	0.2770	0.2769	0.2767	0.2766	0.2765

Figure 44. Imported data table

8.1.2) Plotting data

It is always important to plot your spectra so you can see if you have any bad sample and if the overall data looks as it should. First, select the data to plot with the mouse and then go to ***Plot→Line*** and leave the drop off menu at "All variables" so it will plot all wavelength range. Be sure you select by rows and not by columns, otherwise when you plot your data you won't plot sample spectra but wavelength values in function of the samples. If you do it right, you should get something like Figure 45 (a), otherwise you will get something like Figure 45 (b).

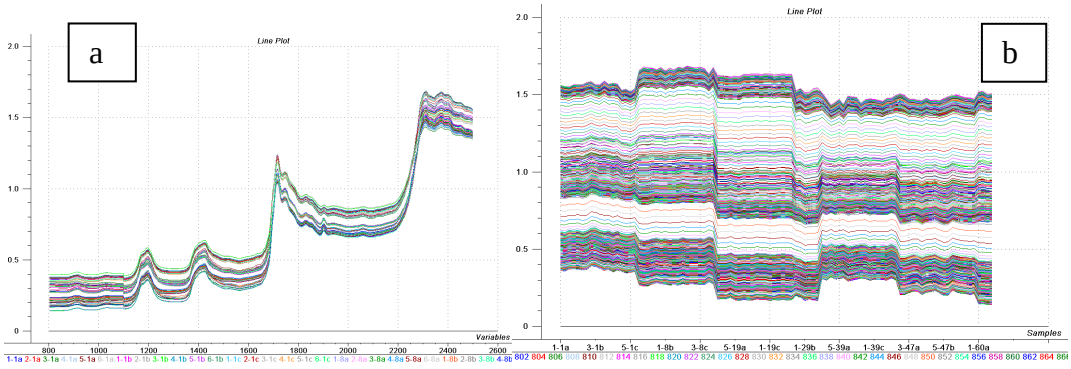


Figure 45. Right spectra (a) and bad plot (b)

One interesting thing to notice in the plot is a kind of cut in the wavelength value 1100. It is not a problem, what it is is a change on the instrument detector: The instrument has two detectors (remember that some detectors work better at certain wavelength ranges) so some type of reading discontinuity is unavoidable. The best thing to do is not consider that small range of 3 wavelengths in the analysis.

8.1.3) Carrying out PCA

Close the plot window to be back to the data table. No select ***Task*** → PCA. You should get a window where you can select several options. First notice there are two tabs, one with ***samples*** and the other with ***variables***. Let's start with samples. The sample set, will be all samples or selected samples (use the 100 samples we have). If you ever want to ignore any sample, you can type them in the ***Keep out of calculation*** section or you can define the samples to work clicking on ***define*** and following the steps.

Let's go to the validation Method. We will use the ***Cross validation*** method. Click on Setup. You will see that you get the chance to select how you want to do your cross validation from the drop off menu (Figure 46). We advise to leave it **random** or if you have few samples (less than 50 for instance) select the ***Full cross validation***. The reason why is that if you have few samples and you split them in subgroups, you will be creating models with such small number of samples that will be really different from the last real model. On the opposite, if you have a lot of samples and you select full cross validation the program will take forever to do the validation. Because we have 100 samples, we can select random and you can either adjust the ***number of segments*** (number of submodels to be created) or ***samples per segment*** (the number of samples included in each submodel). Let's adjust that last option so we have 20 samples per segment so this leaves the number of segments equal to 5. Click ok.

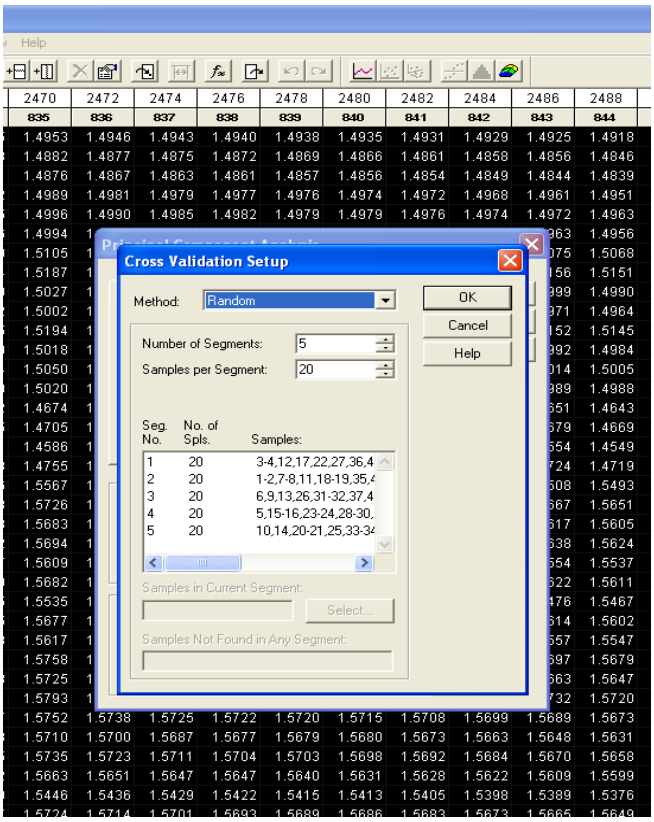


Figure 46. Cross validation Setup

Once on the main window, check the model size. If you leave it in Full and 20 PCs, the software will calculate up to 20 PCs for you. We don't want so many for now because it would take longer, select 10 PCs. Once the window looks like Figure 47, select the variables tab, where you can select the wavelengths you want to include in the analysis. Remember we saw that discontinuity due to a change of detectors and we want to leave those wavelengths out. To make it much easier, we will include in the analysis wavelengths above 1104 nm and ignore shorter wavelengths. To do that, on the ***variable*** tab, check the ***Variable Set*** section. Right now, all variables are selected. Click on ***define*** to enter the desired range. The ***Set Editor*** will pop out. Right now it is empty because you have not defined any range, so click ***Add***. The New Variable Set window will show: enter a name (for instance, short spectra), Data type: Spectra (it does not matter but it will help you remember) and then fill the interval section. To do so, you can either click on select and add it through dragging with the mouse, or you can enter the range using a dash.

Note that you don't enter wavelengts in the range but the column number. The wavelength 1104 is located in column 152, so the range would be 152 – 849 (Figure 48)

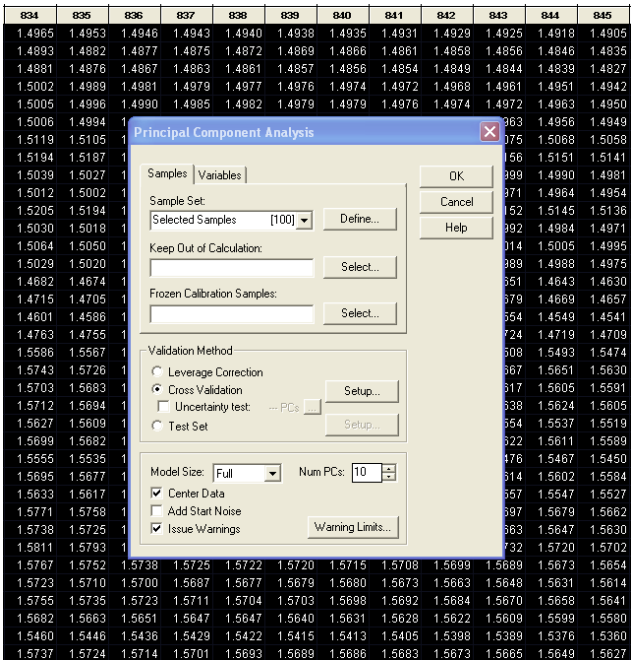


Figure 47. Setting the sample tab options

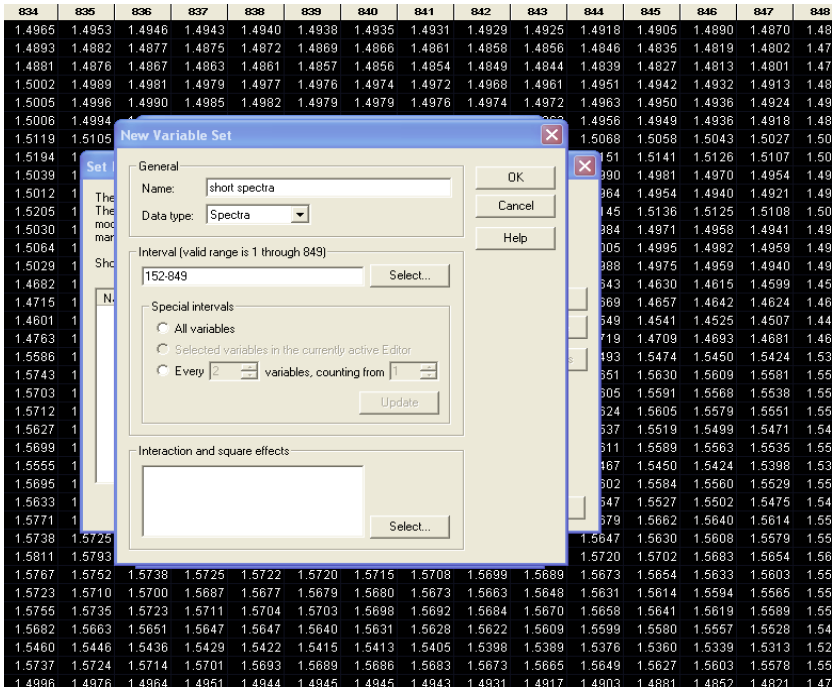


Figure 48. Adding a wavelength range

Everything is ready. Click ok to the ***New Variable Set*** window, ok to the **Set Editor**, and ok to the **Principal Component Analysis** window. The program should start calculating. You will see in the progress window how it displays blue bars respresenting the variance in each component. As we previously explained the first PCs explain most of the variance, so you should see a larger bar in the first PC and then the bars are getting shorter (sometime the variance is so small that the rest of PCs do not get any bar). This process is done as many times as submodels (we had 5 submodels). If you ever see that the variance bars do not decrease but increase, this is a sign that the specific submodel has some trouble: it is very easy that during the random selection of samples one of the submodels get samples from a smaller range and when it is validated it fails. Sometimes, it may also indicate we have outliers.

8.1.4) Checking the results

Once the calculation is finished, you can click on ***View*** to check the results. Four plots are displayed by default, although there are other plots you could check.

Score Plots. The most interesting plot is the scores plot (Figure 49), there is where you will see if there is any sample relationship and clusters. We have clear clustering in our data, like 6 clusters. Knowing the data in advance, we know that this is due to 6 different pipe sizes. Now you can see how PCA can be used to find similarities among samples. On the bottom, it gives you the percentage of variability that each PC represent (first PC represents the 79%, the second PC. A 21%). You can check the scores from other PCs clicking on that plot and playing with the green arrows on the right top. For instance, if you represent the scores on PC2 vs scores on PC3, you will see there are 4 samples pretty far from the rest (Figure 50). PC3 represents such small variability that the program displays 0% .

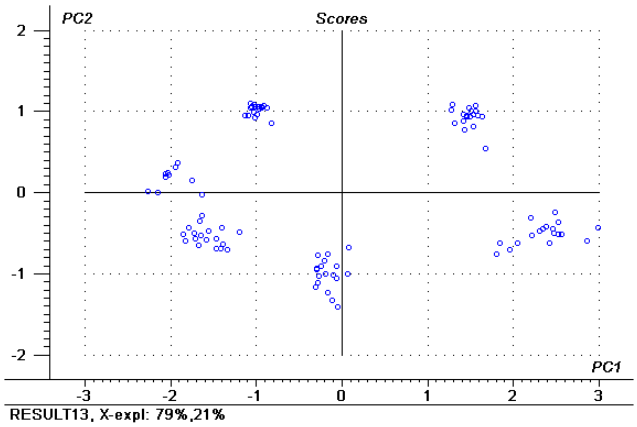


Figure 49. Score plot of PC1 vs PC2

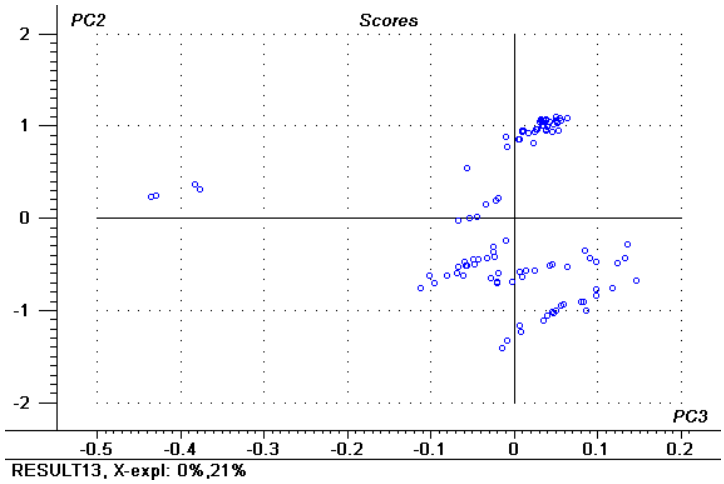


Figure 50. Score plot of PC3 vs PC2

Influence Plot. You can check for problematic samples in the influence plot, where sample residual is plotted versus its leverage (Figure 51). We can check there for samples that have been difficult to model during the PCA process, and samples that have a big influence on it. In our exercise, those 4 samples we saw in figure 50, are the ones with highest residuals so they probably were not well measured. If you would like to calculate the model without them, you would select them using the right icon on the top left and select them with the mouse, and then go again to ***tasks→ recalculate without***

marked. There are a couple of samples with high leverage: those are shown to be a little different than the others and have a higher influence on the model, but they have low residuals so are important samples to keep.

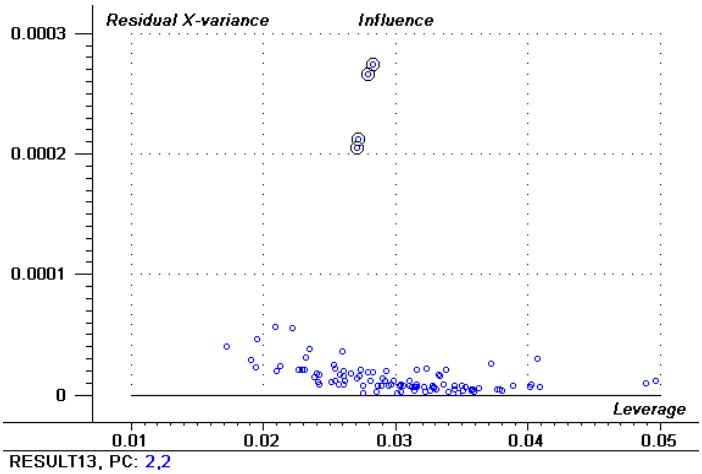


Figure 51. Influence plot showing four samples with high residual

Explained/Residual Variance Plot. This plots helps you determine how many PCs you want to use and keep. The software stops in base of a conservative significance test, so you can use the suggested number of PCs (in this exercise is two, you can see this on the bottom of the influence plot) or you can use your own number using the green arrows on the top right of the window. The explained variance plot though it is a good visual way to see the variance represented by each PC. Personally, I found that when displayed as residual variance is found much useful as you can see the elbow that helps determine when to stop adding PCs in your model. In case you don't have the residual variance display by default (Figure 52), the icon is the last one on the top of the window.

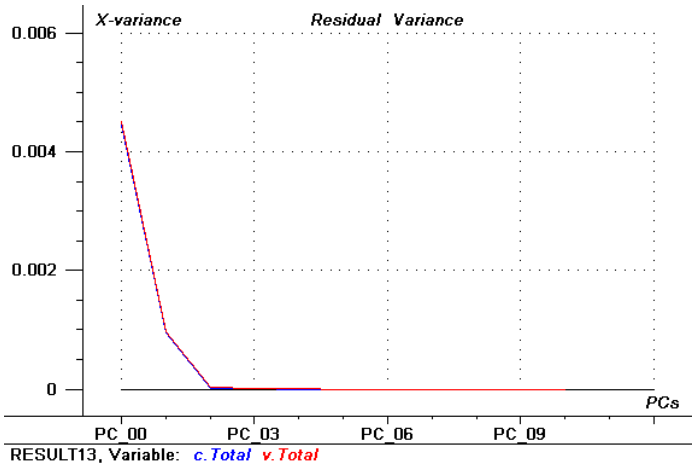


Figure 52. Residual Variance Plot, showing that with 2 PCs almost all data variance can be explained

X-loadings plot. This plot is helpful to determine which wavelengths have more relevance in each PC. For instance, on the first PC (Figure 53) wavelengths above 2300 nm do not seem to have much relevance, while from 1650 to 1700 nm and around 1200 nm the wavelengths seem to have the highest relevance. You can click on the green arrows on the top to check other PCS. The second PC should have a high influence from the highest wavelengths and again, from wavelengths around 1700 nm.

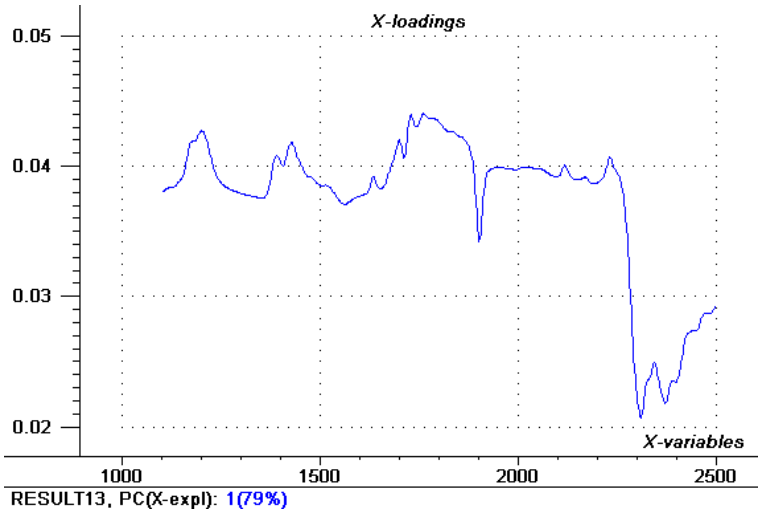


Figure 53. Loading values from all the wavelengths on the first PC

8.1.5) Saving the model and applying it

Once you are happy with it (after removing samples and recalculate it if required) you only need to save it going to ***File → save as***. It automatically lets you browse for the directory and enter a name, and the file is saved as a calibration result (check how under the file name it says "save as type: PCA"). The saved model has a .M extension. You can now try to apply the model. Remember that to apply the model, you need to have data from the same instrument and remember the work conditions: we worked with wavelengths above 1104 nm.

Just to try, close the results window and be back to the data table. We will now apply our model to the same data to practice. Go to task → Project samples (what a PCA model does is project the new spectra to the new variables, PCs. Our model has 2 PCs). By default, all the fields form the Project samples window should be fine, because we used the same data to create the model. If the data were new, you should be sure that you go to the variable tab and you select the same wavelength range. Otherwise, you won't be able to use the model (The Unscrambler won't even find it). Right now, we are fine, so you can go to the **Model Name** section and click on ***Find*** to browse for the model. Once you find it, click OK. You should get the same plots we got before, because it is the same data, but being new spectra you should be able to check the plots to search for answers.

8.2) Developing a basic PLS calibration

8.2.1) Checking data

You will need the data file "*SoybeanOil.OOD*". It is a The Unscrambler file so you do not need to import it, you can directly open it. The file is obtained from a transmittance instrument that measures from 850 nm to 1048 nm at 2 nm increments. There are two columns with reference values corresponding to two compounds: Protein and oil. We will carry out a calibration for oil, and you can later carry out a calibration for protein for further practice.

First, let's check the sample spectra. Select all samples (rows) and click on plot → line. Now be careful, for variable set it says "All variables". You do not want to plot all variables, because remember you have two columns that are not spectral data. So you will have to define a new variable set (click on define, add a new set, and select the right range: from column 3 to the last column). You should have 100 wavelengths or variables to plot. So far, we cannot see any bad sample (Figure 54)

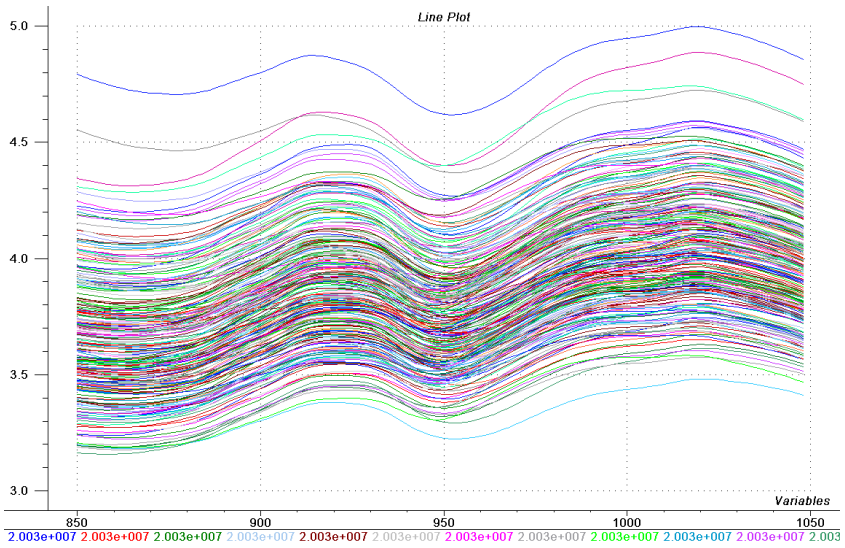


Figure 54. Soybean absorbance spectra from a transmittance instrument

Now, let's check the reference data. Select the column with oil values. Now go to Plot → Histogram. This allows checking for the distribution. Ideally, we would like a uniform distribution, but the common distribution in a population is the normal distribution. We can see this in our histogram (Figure 55), the statistics on the right left can be useful, especially the number of elements or samples, (skewness and kurtossis can be used to check the distribution, but for our purposes, a fast visual inspection is enough), the mean, and the standard deviation. We have a higher proportion of samples with oil ranging from 19.1 to 19.9% approximately. We don't have a big amount of samples, but we could remove 10 samples from that range to make the distribution more uniform. Closing the histogram, go to ***Modify → Sort samples***. We will sort the samples according to their oil content; all samples sorted by ***Values***. In the Keys section of the sort samples

window, set the first variable at 2 (oil is set on the second column) and second variable as 2 as well. Click ok. Now we will find the samples that have oil content higher than 19.1 and we will delete one sample every 3 (delete 59, 62…) up to the samples that have higher content than 20.1%. If you check the histogram again, you will see that looks a little bit better although it is common to have a smaller amount of samples with high and low oil content (having a perfect uniform distribution would mean to have very few samples!)

This is a fast way to do it and maybe not the most advisable as we may be removing interesting samples. There are a couple of sample selection algorithms that take in account sample spectra as well, but they are not very extended or they are part of propiertary software. Right now, it is a valid enough approach.

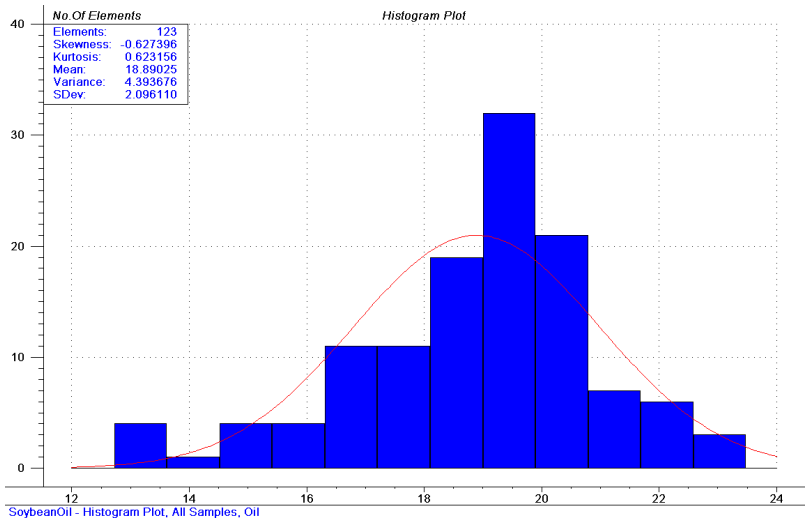


Figure 55. Distribution of oil content (histogram) before deleting samples

8.2.2) Calibration process and results

Now we are ready to start the PLS calibration. Go to Task → Regression . The regression window will open. Right on the top, there are the possible regression methods in The Unscrambler, we will leave it in PLS1 (PLS for 1 compound). Similarly to the last PCA exercise, we will check ***cross validation*** as the ***Validation Method*** . We will pick this time full cross-validation even if it takes a little longer (it means the program will

create over 100 submodels). Click ok, and then from the regression window set the number of PCs to calculate to 12 (right now, 12 is enough).

Go to the X-variable tab and set the right interval. That is, all the wavelength range (column 3 to the end). Do the same with the Y-variable, which is the compound to be measured. For this first example is oil, so it is just the second column. When you have the set-up ready (Figure 56) click OK. We will check the plots we obtain.

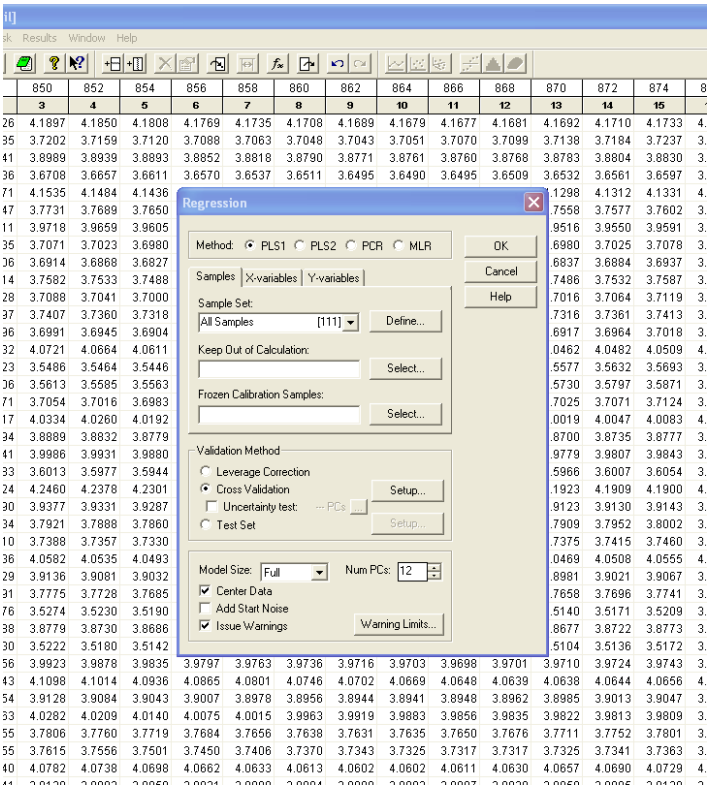


Figure 56. regression window correctly filled

Scores Plot. Like in PCA, the score plot can be check in PLS, but remember that we are not exactly checking PCs but Latent Variables although the program gives them the same name, those have been calculated a slightly different than PCs, taking in account the reference values (oil content). The plot is not one of the most important this time, but it may still be useful to see any strange data clustering. If you see clustering in PLS, be careful. You should apply preprocessing methods and see of those can be eliminated,

different clusters may lead to non-linearities and may mean that you have completely different populations in your calibration set (could be an indication of different reference methods used for instance…you should check why).

Residual/Explained Validation Variance Plot. This is again the plot you have to check to know how many LVs should be included in your model (check the PCA exercise). The software again is giving its suggestion (8 PCs, you can see it on the bottom of the two remaining plots). Eight may be a good choice, and you could even add another one (check the new results with 9 PCs clicking on any of the plots and using the green arrows on the top).

Regrescion Coefficients (B). Those are the real results of your calibration model (those are the coefficients you need to multiply to each corresponding wavelenghts from a new spectra to obtain a prediction of oil content), displayed graphically (Figure 57). This plot looks like a spectrum, and it is desirable it looks smooth. On the X axis there are the wavelenghts and on the Y axis there are no units as they are coefficients. On the bottom of the plot, there is the offset value. You may need to check the numerical value of the coefficients. To do so, right click on the plot, go to ***View → Numerical***. The results are displayed for the model with selected PCs (by default, using the 8PCs) (Figure 58). You will note you have fewer values than wavelengths, that's because the wavelengths that got a coefficient equal to 0 are not displayed. You can copy them by right clicking and paste them in word or excel.

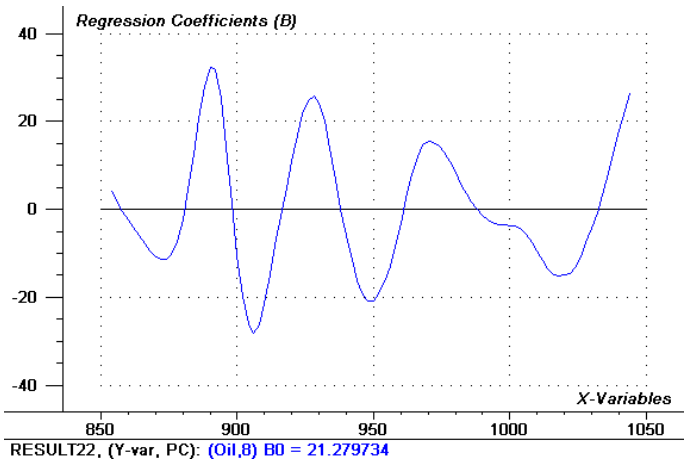


Figure 57. Regression Coefficients plot

So just to check it is clear, the way that those coefficients are utilized in future predictions would be:

Predicted Oil content (%) = B0 + coefficient1*absorbance1 + coefficient2*absorbance2 +…

Predicted vs Measured Plot. This is the plot you should use to report your calibration and your calibration validation by cross-validation. By default, it is showing two values for each of the parameters in the box. The most interesting ones are RMSE and R-square. If you click on the plot and then you unclick on the icon on the top and to the right that says **Cal**, the plot will only display the cross-validation results and you will get more statistics. Those are the approximate results you should get doing so (Figure 58 a and b) I say approximate because since our cross-validation is random, your subsets and mine will be different and will lead to slightly different results.

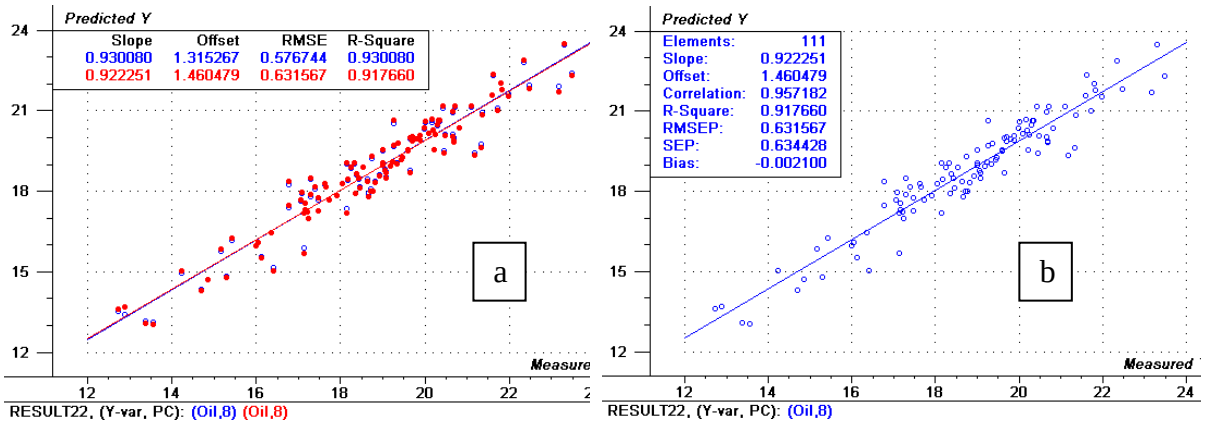


Figure 58. Predicted vs measured of both calibration and validation (a) and just for validation (resulting from the cross-validation submodels) (b)

Getting the results from just the validation, allows obtaining more statistics. Here you can find the statistics we previously mentioned: Bias, SEP, RMSEP, R-square. Be sure you really uncheck Cal instead of Val, otherwise you should notice that instead of

displaying SEP, SEC (standard error of calibration) is displayed. You should see that we have a good calibration according to the results.

Influence Plot. This plot is not shown by default, so you should click on any of the plots you do not need and then go to Plots→Residuals and check influence plot. You have other plots you may want to see, although with the ones we already saw most of the relevant information have been checked. We have already seen the Influence plot in the PCA analysis and we have seen that it is very useful to detect outliers. In our example, we do not have clear outliers, but we have a set of samples that show high residuals (Figure 59). You can try to delete them (first, click on the right icon on the top and select the samples) and recalculate the model (***task→recalculate without marked***). You should get a significant improvement (my SEP drop to 0.5% from the initial 0.63%).

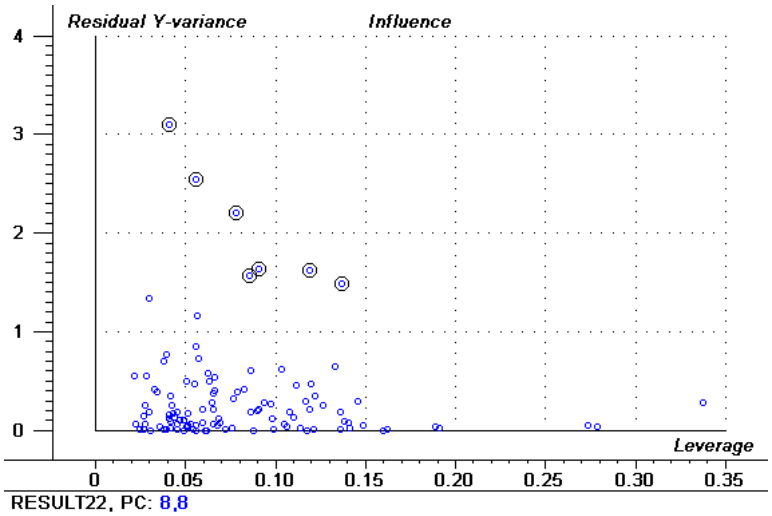


Figure 59. Influence plot with suggested samples to delete

8.2.3) Saving and applying the model

Similar to the PCA example, go to ***file→ save as***, and save your regression model. This time, we will try to use this model to predict from new data. The trick is that the data we will use (called "SoybeanOilVal.OOD") comes from another instrument but from the same brand. We will see what happens.

One you have your model saved, now close your data table too and open the validation file ("SoybeanOilVal.OOD"). Before you predict, you should first select the variables (the scans). If you don't do that, The Unscrambler will think that all your data tables are scans only, and when you will try to find the model it won't let you use it (in fact, you won't even have the chance to select it) because it has been created with 100 variables and not 102. Be careful with that, this is one of the tricky things with The Unscrambler. To select the variables you can go to ***Edit→Select variables*** and then Define a new set. You only have to do what you did before with your first data, enter the right range where you have the spectra (column 3 to 102). You could also select them manually (highlight them), but this way is much easier.

One you have the spectra selected, go to ***task → predict*** and browse for the model. (in Model Name section, click on Find…). Now go to the ***Y-reference*** Tab and click "***Include Y-Reference***" (Figure 60). We can do that because we have the reference values of our data and we can validate the model again; if we did not and we only need the rpedictions, then leave it like that. Then you have to define the column that has the reference data (the second column again). You can click Ok and The Unscrambler will predict for you.

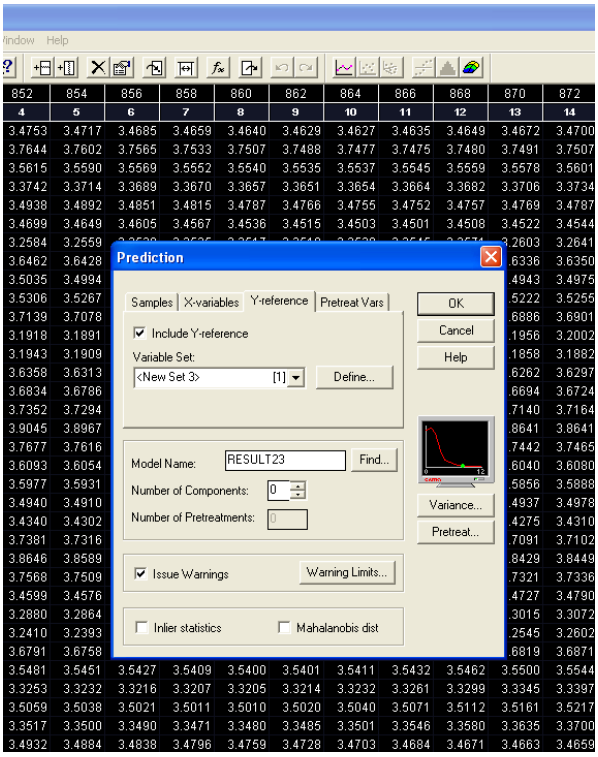


Figure 60. Setting the Reference in the prediction box

By default, the results will be displayed as box plots and numerically (Figure 61), but we want to get the statistics (SEP, R-Square…). So you can go to ***Plots → prediction*** and then select ***Predicted vs Reference***. You should see how we did not get bad results at all (Figure 62): bias increased significantly (as expected, the accuracy will get slightly worse) but SEP just increased a little. This should give you a taste of what would come next: how to standardize or transfer a calibration from one instrument to another.

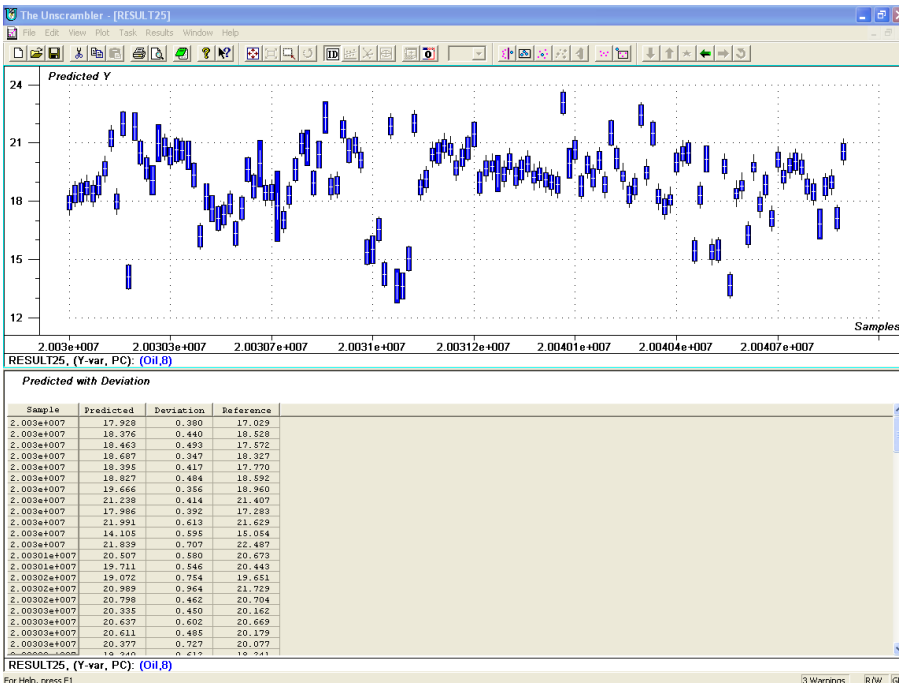


Figure 61. Default display for prediction results

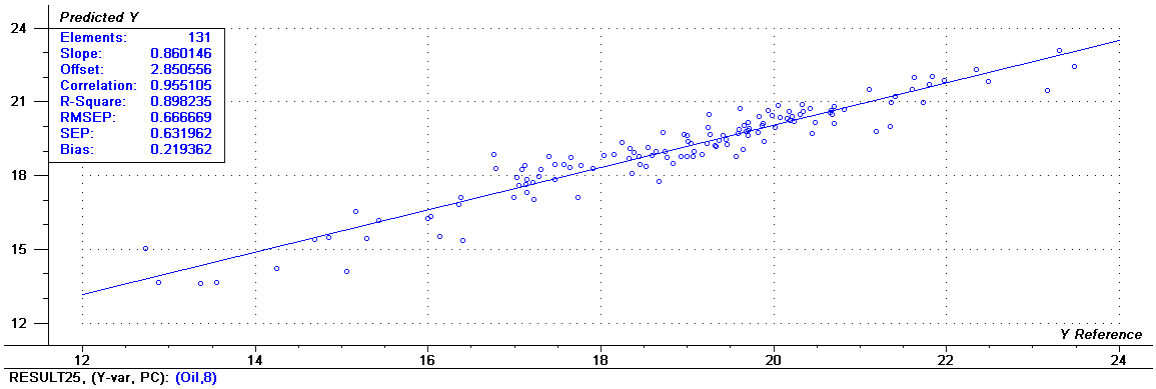


Figure 62. Prediction results: predicted vs reference

A suggestion is you try now to carry out a calibration with protein and then try to predict from the validation file. You can also try to apply preprocessing methods (go to Modify→ transform and select a few), is any of them giving better models for both cross-validation and "independent" validation (using SoybeanOilVal file and predict)?

9. MATLAB EXAMPLES AND CODES

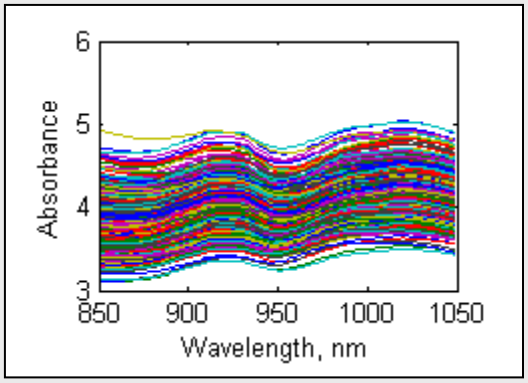
MATLAB programming environment in combination with toolboxes, individual custom-written functions, and scripts is an extremely versatile and powerful software package for analysis of NIR data. Effective utilization of this software requires general programming skills, knowledge of MATLAB programming language, and strong familiarity with the available tools and features. However, MATLAB has a very useful help material for its functions and toolboxes and some step-by-step examples. We provide in this section some useful exercises and example codes to help users start using matlab environment and PLS_toolbox for NIR analysis. The examples were originally created with Matlab 7.0.1 (R14), LS_SVMlab v.1.5, and PLS_Toolbox 3.0.1 but have been further tested with Matlab 7.0.4 and PLS_Toolbox 3.5.4. For more information on MATLAB functionality and features refer to technical documentation (www.mathworks.com), help files, and demos. All data files used in the examples are located on the Grainbin file server (\\grainbin\Users\Shared\NIR Primer).

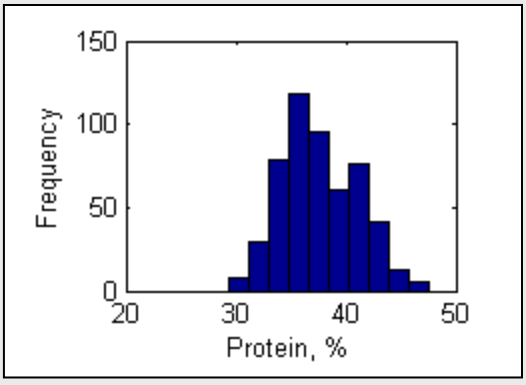
9. 1) Importing XLS files

Task: Import spectral and reference data from file “matlabEx8p1DataFile.xls” into MATLAB.

```
>> % First, make sure the data file is located in the current
>> % working directory. Otherwise, instead of just file name,
>> % use full path, for example ‘C:\Matlab\work\filename.xls’.
>>
>> [data1]=xlsread('matlabEx8p1DataFile.xls','Sheet1');
>> [data2]=xlsread('matlabEx8p1DataFile.xls','Sheet2');
>> [data3]=xlsread('matlabEx8p1DataFile.xls','Sheet3');
>>
>> whos
  Name        Size                    Bytes  Class

  data1     102x251                  204816  double array
  data2     102x251                  204816  double array
  data3     102x27                    22032  double array

Grand total is 53958 elements using 431664 bytes

>> % Save sample IDs.
>> IDs=[data1(1,2:end),data2(1,2:end),data3(1,2:end)];
>>
>> % Save wavelengths.
>> wLens=[data1(2:end-1,1)];
```

```
>> % Merge spectral data from data1, data2, and data3 into one
>> % matrix spectra.
>> spectra=[data1(2:end-1,2:end),data2(2:end-1,2:end),...
data3(2:end-1,2:end)];
>>
>> % Merge reference data into one vector prot.
>> prot=[data1(end,2:end),data2(end,2:end),data3(end,2:end)];
>>
>> % Delete unnecessary variables.
>> clear data1 data2 data3;
>>
>> whos
  Name          Size                    Bytes  Class

  IDs           1x526                    4208  double array
  prot          1x526                    4208  double array
  spectra     100x526                  420800  double array
  wLens       100x1                       800  double array

Grand total is 53752 elements using 430016 bytes

>> % Since it is more customary to store variables in columns and
>> % samples in rows, transpose matrix spectra and vector prot.
>> spectra=spectra'; prot=prot'; IDs=IDs';
>> whos
  Name          Size                    Bytes  Class

  IDs         526x1                      4208  double array
  prot        526x1                      4208  double array
  spectra     526x100                  420800  double array
  wLens       100x1                       800  double array

Grand total is 53752 elements using 430016 bytes

>> % For visual inspection of the data, plot spectra and protein.
>> plot(wLens, spectra);
>> xlabel('Wavelength, nm'); ylabel('Absorbance');
>>
>> figure, hist(prot);
>> xlabel('Protein, %'); ylabel('Frequency');
>>
```

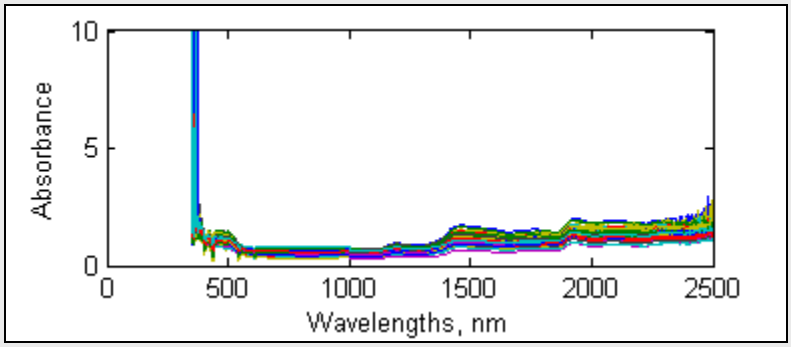




9. 2) Importing TXT files

Task: Import spectral data from file “matlabEx8p2DataFile.txt” into MATLAB.

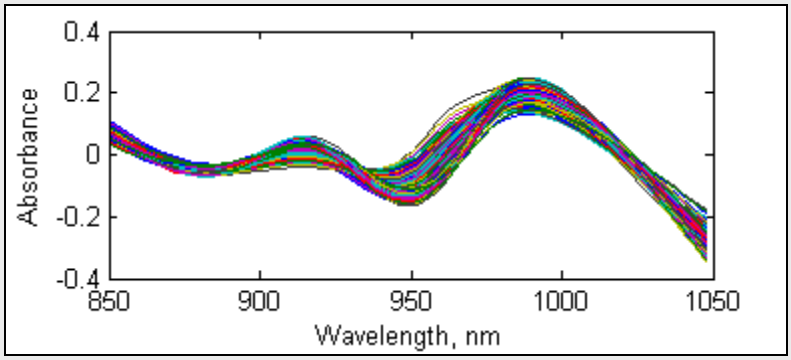
```
>> allData=importdata('matlabEx8p2DataFile.txt');
>>
>> % Variable allData is a structure with three fields:
>> allData

allData =

          data: [2151x61 double]
      textdata: {1x61 cell}
    colheaders: {1x61 cell}

>> % Save spectral data.
>> spectra=allData.data(:,2:end);
>>
>> % Save wavelengths.
>> wLens=allData.data(:,1);
>>
>> % Save sample IDs.
>> IDs=allData.colheaders(2:end);
>>
>> % Delete allData structure and transpose IDs and spectra.
>> clear allData;
>> IDs=IDs'; spectra=spectra';
>> whos
  Name          Size                     Bytes  Class

  IDs          60x1                     11040  cell array
  spectra      60x2151                1032480  double array
  wLens      2151x1                     17208  double array

Grand total is 134991 elements using 1060728 bytes

>> % Plot spectra.
>> plot(wLens,spectra);
>> xlabel('Wavelengths, nm'); ylabel('Absorbance');
>>
```



9. 3) Importing SPC files

Task: Import spectral data from file “matlabEx8p3DataFile.spc” into MATLAB.

```
>> [spectra,wLens]=spcreadr('matlabEx8p3DataFile.spc');
>>
>> whos
  Name          Size                     Bytes  Class

  spectra     625x100                   500000  double array
  wLens         1x100                      800  double array

Grand total is 62600 elements using 500800 bytes

>> wLens=wLens';
>> plot(wLens,spectra);
>> xlabel('Wavelength, nm'); ylabel('Absorbance');
```



9. 4) Saving MATLAB data in TXT and CSV files

Task: Save selected variables from MATLAB workspace as TXT and CSV files.

```
>> load matlabEx8p4DataFile.mat
>> save('spectra.txt', 'spectra', '-ASCII');
>> save('protein.txt', 'prot', '-ASCII');
>>
>> csvwrite('spectra.csv', spectra);
>> csvwrite('protein.csv', prot);
>>
```

9. 5) Storing spectral and reference data in MAT files

Task: Store spectral and reference data in MAT file for subsequent use in MATLAB and The Unscrambler.

Note: Advantage of storing data in MATLAB file format vs. MS Excel or MS Access is that MAT database does not have limitation on number of columns (Excel and Access tables are limited to 254 columns).

```
>> % Resultant MAT file should contain following variables:
>> %     fileInfo        - general info about data; data class:
>> %                       cell array.
>> %     refData         - reference chemistry data; data class:
>> %                       double-precision vector or matrix.
>> %     refDataLabels   - reference data labels; data class: matrix
>> %                       of characters.
>> %     sampleIDs       - sample IDs; data class: numerical
>> %                       double-precision column vector.
>> %     sampleIDsChar   - sample IDs for reading by
>> %                       The Unscrambler; data class: matrix of
>> %                       characters.
>> %     spectra         - spectral data (samples in rows, variables
>> %                       in columns); data class: numerical
>> %                       double-precision matrix.
>> %     wLens           - wavelengths in nm; data class: numerical
>> %                       double-precision column or row vector.
>> %     wLensChar       - wavelengths in nm for reading by The
>> %                       Unscrambler; data class: matrix of
>> %                       characters.
>>
>> % First, perform data import steps from the example
>> % “8.1. Importing XLS files”.
>>
>> [data1]=xlsread('matlabEx8p1DataFile.xls','Sheet1');
[data2]=xlsread('matlabEx8p1DataFile.xls','Sheet2');
[data3]=xlsread('matlabEx8p1DataFile.xls','Sheet3');
IDs=[data1(1,2:end),data2(1,2:end),data3(1,2:end)];
wLens=[data1(2:end-1,1)];
spectra=[data1(2:end-1,2:end),data2(2:end-1,2:end),...
data3(2:end-1,2:end)];
prot=[data1(end,2:end),data2(end,2:end),data3(end,2:end)];
clear data1 data2 data3;
spectra=spectra'; prot=prot';IDs=IDs';
>>
>> whos
  Name          Size                    Bytes  Class

  IDs           1x526                    4208  double array
  prot        526x1                      4208  double array
  spectra     526x100                  420800  double array
  wLens       100x1                       800  double array

Grand total is 53752 elements using 430016 bytes
```

```
>> % Even though wLens variable is a column header and it would be
>> % logical to transpose it into a row vector, we are not going to
>> % do it.  Here is why: conversion of a numerical row vector into
>> % character variable will result in one long string of characters
>> % which will not be understood by The Unscrambler as wavelength
>> % variable names.
>>
>> fileInfo={'Commodity: soybeans'; 'Instrument: FOSS Infratec';...
'Author: Name'; 'Date: xx/xx/xxxx'; 'Comments: ...'};
>>
>> refData=prot;
>> refDataLabels=['Protein'];
>>
>> sampleIDs=IDs;
>> sampleIDsChar=num2str(sampleIDs);
>>
>> wLensChar=num2str(wLens);
>> clear IDs prot;
>>
>> whos
  Name                Size                    Bytes  Class

  fileInfo            4x1                       372  cell array
  refData           526x1                      4208  double array
  refDataLabels       1x7                        14  char array
  sampleIDs         526x1                      4208  double array
  sampleIDsChar     526x8                      8416  char array
  spectra           526x100                  420800  double array
  wLens             100x1                       800  double array
  wLensChar         100x4                       800  char array

Grand total is 64212 elements using 451168 bytes

>> % Before saving data, make sure that MATLAB is set to save MAT
>> % files in v.6 format (The Unscrambler v.9.1 does not read MATLAB
>> % v.7 files). In MATLAB, go to File/Preferences/General/MAT-Files
>> % and select “Ensure backward compatibility (-v6)”.
>>
>> % Save file in the current working directory.
>> save 'SbFossInfratecProteinCalData.mat';
>>
```

9. 6) Opening MAT files in The Unscrambler

Task: Import MAT data file into The Unscrambler v. 9.1.

Solution:

a) Start The Unscrambler.

b) First, import spectral data: go to File/Import/Matlab and locate MAT file to be imported (see Figure 8.1). Click Import.

c) In the next window, select spectra in the Data field, sampleIDsChar in Sample names, and wLensChar in Variable names (Figure 63). Click OK.

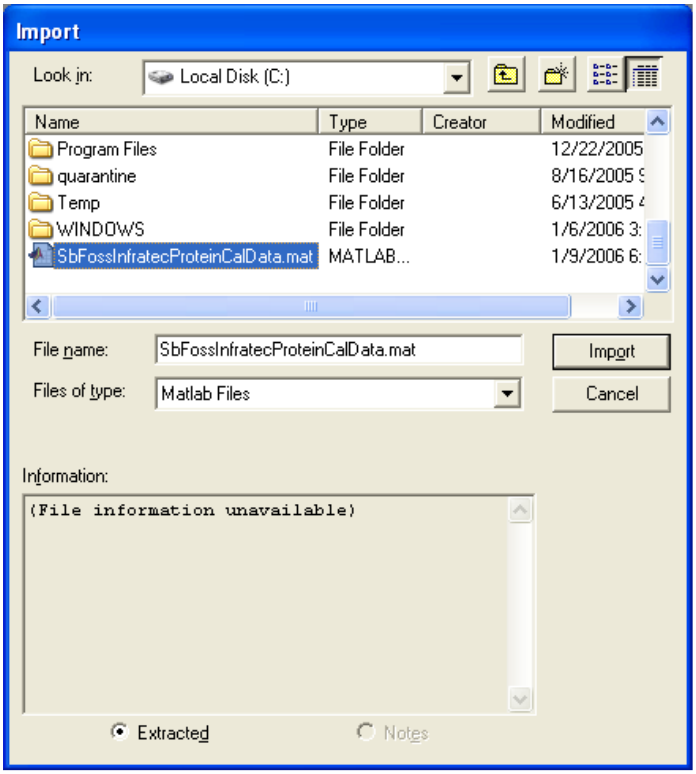


Figure 63. Locating MAT file to be imported.

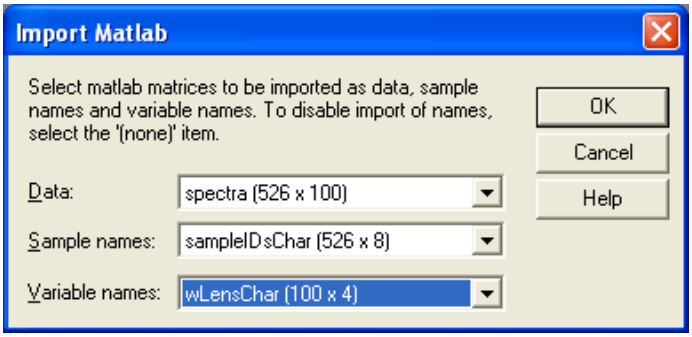


Figure 64. Selecting MAT file variables for importing spectral data.

d) Insert an empty 1st column that in the next few steps will be filled with the reference values. Go to Edit/Insert/Variable.

e) Import reference data: go to File/Import/Matlab and select “Current data table (from origin)” (Figure 65). Click OK.

f) Locate the same MAT file and click Import.

g) In the next dialog window, select reference data vector (refData), sample names (sampleIDsChar), and reference variable name (refDataLabels) as shown in Figure 66. To finish data import, click OK. The result is shown in Figure 67.

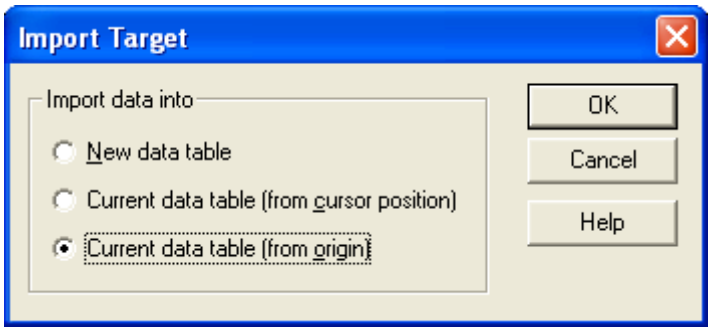


Figure 65. Selecting destination for reference data.

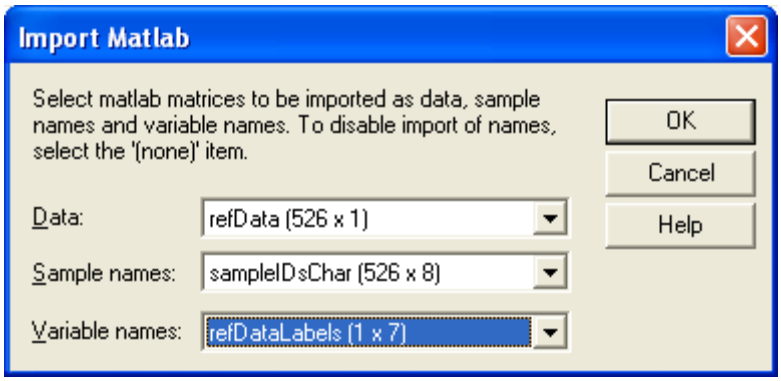
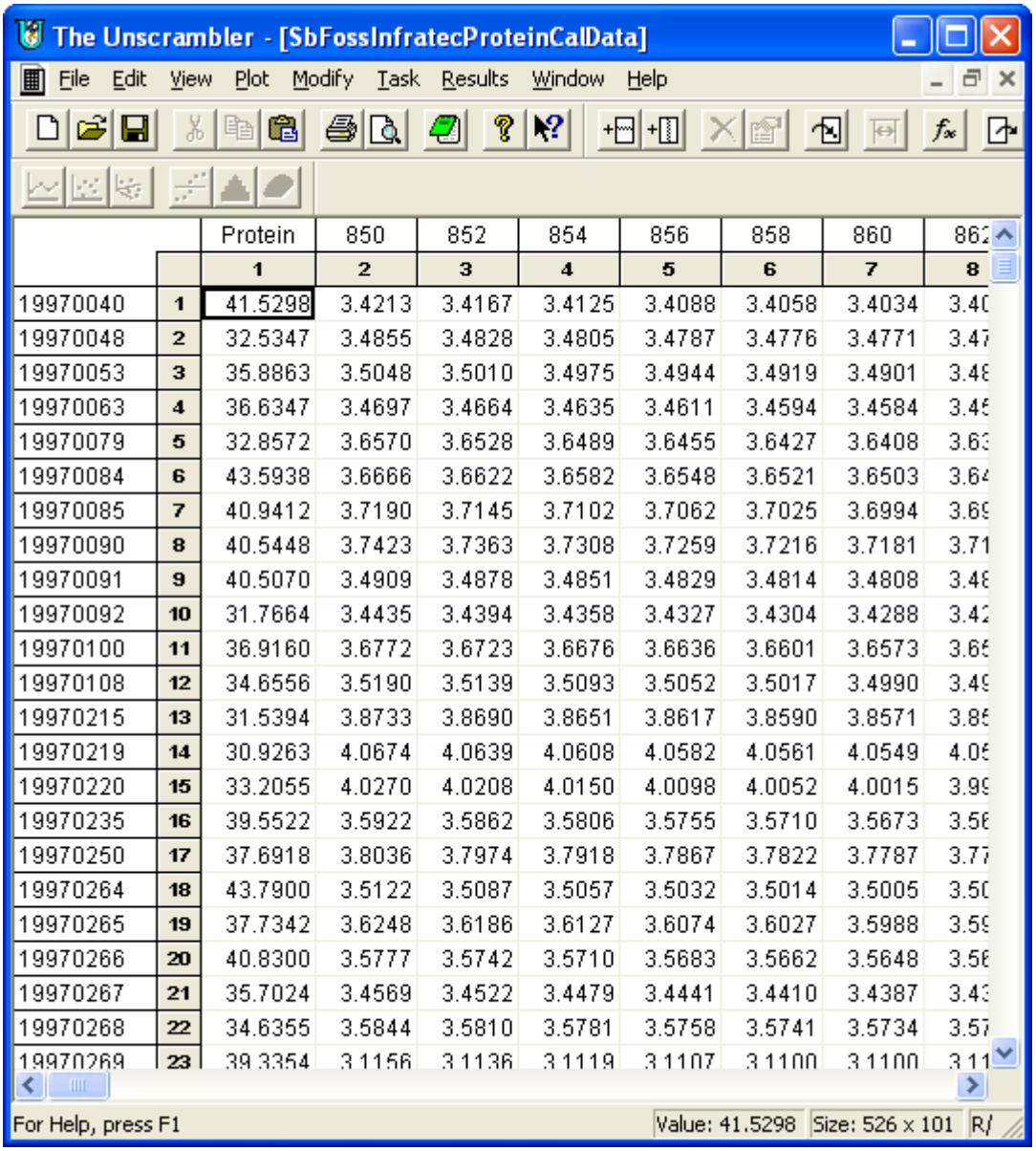


Figure 66. Selecting MAT file variables for importing reference data.



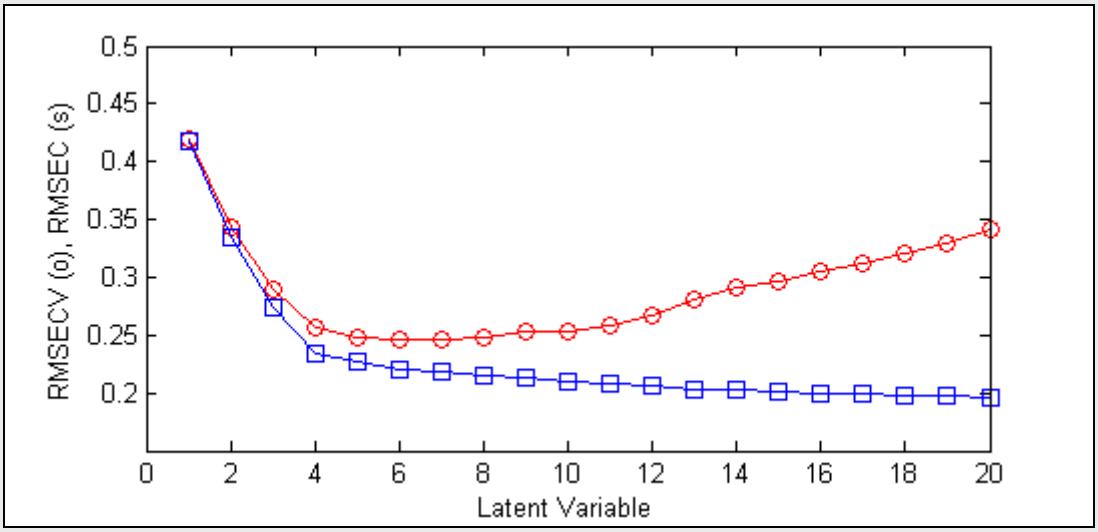
		Protein	850	852	854	856	858	860	862
		1	2	3	4	5	6	7	8
19970040	1	41.5298	3.4213	3.4167	3.4125	3.4088	3.4058	3.4034	3.40
19970048	2	32.5347	3.4855	3.4828	3.4805	3.4787	3.4776	3.4771	3.47
19970053	3	35.8863	3.5048	3.5010	3.4975	3.4944	3.4919	3.4901	3.48
19970063	4	36.6347	3.4697	3.4664	3.4635	3.4611	3.4594	3.4584	3.45
19970079	5	32.8572	3.6570	3.6528	3.6489	3.6455	3.6427	3.6408	3.63
19970084	6	43.5938	3.6666	3.6622	3.6582	3.6548	3.6521	3.6503	3.64
19970085	7	40.9412	3.7190	3.7145	3.7102	3.7062	3.7025	3.6994	3.69
19970090	8	40.5448	3.7423	3.7363	3.7308	3.7259	3.7216	3.7181	3.71
19970091	9	40.5070	3.4909	3.4878	3.4851	3.4829	3.4814	3.4808	3.48
19970092	10	31.7664	3.4435	3.4394	3.4358	3.4327	3.4304	3.4288	3.42
19970100	11	36.9160	3.6772	3.6723	3.6676	3.6636	3.6601	3.6573	3.65
19970108	12	34.6556	3.5190	3.5139	3.5093	3.5052	3.5017	3.4990	3.49
19970215	13	31.5394	3.8733	3.8690	3.8651	3.8617	3.8590	3.8571	3.85
19970219	14	30.9263	4.0674	4.0639	4.0608	4.0582	4.0561	4.0549	4.05
19970220	15	33.2055	4.0270	4.0208	4.0150	4.0098	4.0052	4.0015	3.99
19970235	16	39.5522	3.5922	3.5862	3.5806	3.5755	3.5710	3.5673	3.56
19970250	17	37.6918	3.8036	3.7974	3.7918	3.7867	3.7822	3.7787	3.77
19970264	18	43.7900	3.5122	3.5087	3.5057	3.5032	3.5014	3.5005	3.50
19970265	19	37.7342	3.6248	3.6186	3.6127	3.6074	3.6027	3.5988	3.59
19970266	20	40.8300	3.5777	3.5742	3.5710	3.5683	3.5662	3.5648	3.56
19970267	21	35.7024	3.4569	3.4522	3.4479	3.4441	3.4410	3.4387	3.43
19970268	22	34.6355	3.5844	3.5810	3.5781	3.5758	3.5741	3.5734	3.57
19970269	23	39.3354	3.1156	3.1136	3.1119	3.1107	3.1100	3.1100	3.11

Figure 67. MAT file imported into The Unscrambler.

9. 7) Calibration using PLS regression

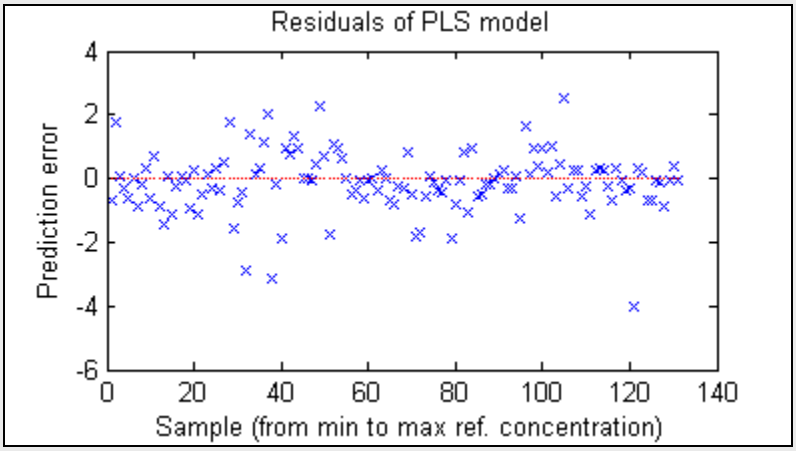
Task: Develop and validate PLS calibration model using data set "SbFossInfratecProteinCalData.mat".

```
>> % Load data file.
>> load SbFossInfratecProteinCalData.mat;
>>
>> % Preprocess spectra (2nd derivative).
>> spectra=savgol(spectra,5,3,2);
>>
>> % Divide data into calibration and validation sets.
>> % Use 75% of the samples for calibration and the other 25%
>> % for testing.
>>
>> inputTrain=spectra([1:4:end,2:4:end,4:4:end],:);
>> targetTrain=refData([1:4:end,2:4:end,4:4:end],:);
>>
>> inputTest=spectra(3:4:end,:);
>> targetTest=refData(3:4:end,:);
>>
>> % Normalize (autoscale) inputs and targets.
>> [inputTrainNorm,meanInputTrainNorm,stdInputTrainNorm]=...
auto(inputTrain);
>> inputTestNorm=scale(inputTest,meanInputTrainNorm,...
stdInputTrainNorm);
>> [targetTrainNorm,meanTargetTrainNorm,stdTargetTrainNorm]=...
auto(targetTrain);
>>
>> % Perform cross-validation to find the best number of
>> % PLS predictors.
>> [press,cumpress,rmsecv,rmsec,cvpred]=...
crossval(inputTrainNorm,targetTrainNorm,'sim',{'con' 5},20);
>>
```

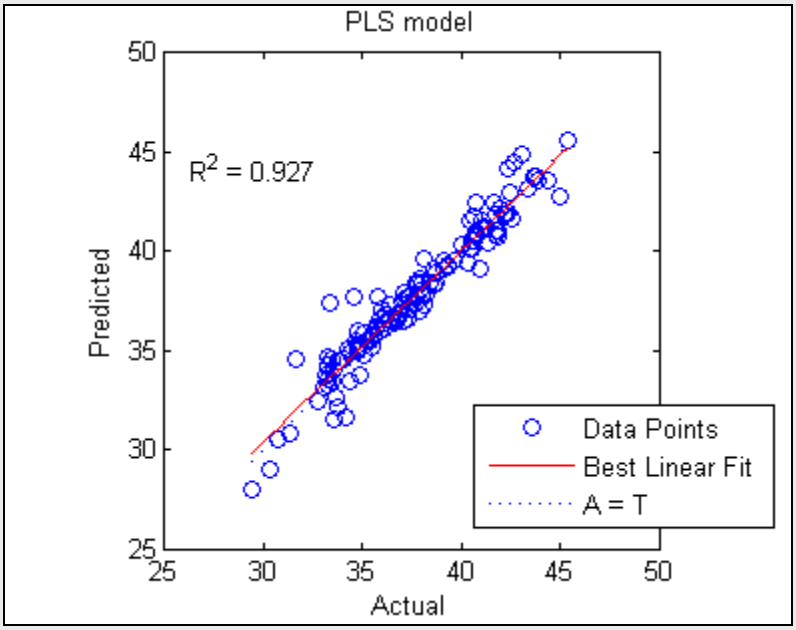


```
>> % From the graph, select number of latent variables corresponding
>> % to the lowest value of RMSECV(0);
>> nLV=6;
>>
>> % Perform PLS regression.
>> plsOptions=pls('options');
>> plsOptions.display='off';
>> plsOptions.plots='none';
>>
>> plsModel=pls(inputTrainNorm,targetTrainNorm,nLV,plsOptions);
>> plsPredictedNorm=pls(inputTestNorm,plsModel,plsOptions);
>>
>> plsPredicted=rescale(plsPredictedNorm.pred{2},...
meanTargetTrainNorm,stdTargetTrainNorm);
>>
>> [plsSlope,plsIntercept,plsR]=postreg(plsPredicted',targetTest');
>> plsBias = sum(targetTest-plsPredicted)./length(targetTest-...
plsPredicted);
>> plsSEP = std(plsPredicted-targetTest);
>> plsRPD = std(targetTest)/plsSEP;
>>
>> PLS.model=plsModel;
>> PLS.numOfLVs=nLV;
>> PLS.residuals=targetTest-plsPredicted;
>> PLS.r=plsR;
>> PLS.rSq=plsR^2;
>> PLS.SEP=plsSEP;
>> PLS.slope=plsSlope;
>> PLS.intercept=plsIntercept;
>> PLS.bias=plsBias;
>> PLS.RPD=plsRPD;
>>
>> close all;
>>
>> % Display results
>> PLS

PLS =

        model: [1x1 struct]
     numOfLVs: 6
    residuals: [131x1 double]
            r: 0.9630
          rSq: 0.9275
          SEP: 0.9404
        slope: 0.9627
    intercept: 1.5166
         bias: -0.1149
          RPD: 3.6790
>>
```

```
>> figure; plot(PLS.residuals,'bx');
>> title('Residuals of PLS model');
>> xlabel('Sample (from min to max ref. concentration)');
>> ylabel('Prediction error');
>> hold on;
>> plot([1:length(PLS.residuals)],0,'r-');
>> hold off;
```




```
>> plsPredictedNorm=pls(inputTestNorm,PLS.model,plsOptions);
>> plsPredicted=rescale(plsPredictedNorm.pred{2},...
meanTargetTrainNorm,stdTargetTrainNorm);
>> figure; postreg(plsPredicted',targetTest');
>> title('PLS model');
>> xlabel('Actual'); ylabel('Predicted');
>>
```



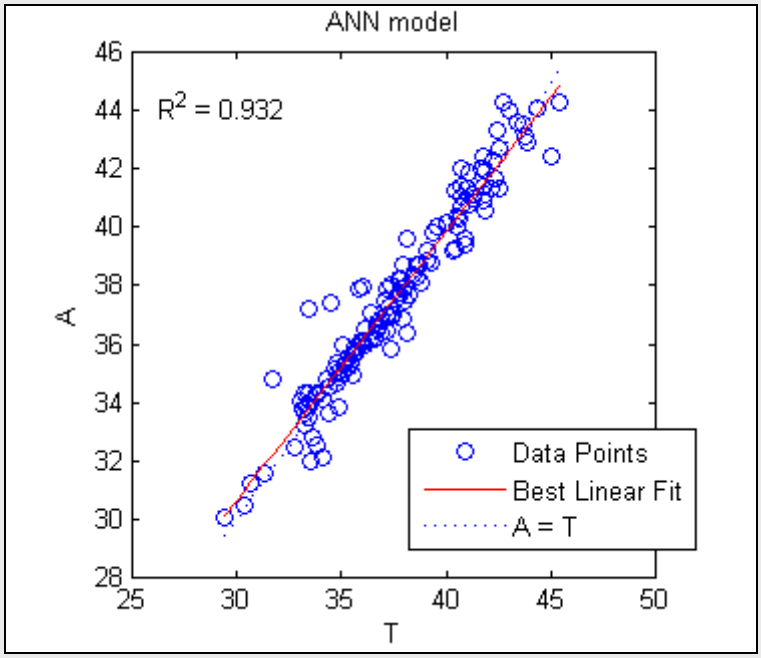
9. 8) Calibration using ANN regression

Task: Develop and validate ANN calibration model using data set "SbFossInfratecProteinCalData.mat".

```
>> % Load data file.
>> load SbFossInfratecProteinCalData.mat;
>>
>> % Preprocess spectra (2nd derivative).
>> spectra=savgol(spectra,5,3,2);
>>
>> % Divide data into calibration and validation sets.
>> % Use 75% of the samples for calibration and 25% for testing.
>> inputTrain=spectra([1:4:end,2:4:end,4:4:end],:);
>> targetTrain=refData([1:4:end,2:4:end,4:4:end],:);
>> inputTest=spectra(3:4:end,:);
>> targetTest=refData(3:4:end,:);
>>
>> % 10% of the calibration data will be used as an "early
>> % stopping set" to prevent overtraining.
>> inputVal=inputTrain(1:10:end,:);
>> targetVal=targetTrain(1:10:end,:);
>>
>> % Transpose matrices for ANN training.
>> inputTrain=inputTrain';
>> inputVal=inputVal';
>> inputTest=inputTest';
>> targetTrain=targetTrain';
>> targetVal=targetVal';
>> targetTest=targetTest';
>>
>> % Normalize spectra.
>> [inputTrainNorm,meanInputTrainNorm,stdInputTrainNorm]=...
prestd(inputTrain);
>> inputValNorm=trastd(inputVal,meanInputTrainNorm,...
stdInputTrainNorm);
>> inputTestNorm=trastd(inputTest,meanInputTrainNorm,...
stdInputTrainNorm);
>>
>> % Reduce number of inputs using PCA compression.
>> minFracOfVar=0.0025;
>> [inputTrainNormTrans,transMat]=prepca(inputTrainNorm,...
minFracOfVar);
>> % prepca finds PCs that are responsible for
>> % (100-100*minFracOfVar)% of total variation;
>> temp=size(inputTrainNormTrans);
>> PCs=temp(1); clear temp;
>>
>> inputValNormTrans=trapca(inputValNorm,transMat);
>> inputTestNormTrans=trapca(inputTestNorm,transMat);
>> val.P=inputValNormTrans;
>> val.T=targetVal;
>>
```

```
>> % Train ANN model.
>> valMse=1000; % initial value of MSE of intermediate testing;
>> numOfNeurons=3; % number of neurons in a hidden layer;
>>
>> for reinit=1:10;
    disp(sprintf('Number of PCs (inputs):       %g',PCs));
    disp(sprintf('Neurons in a hidden layer:    %g',numOfNeurons));
    disp(sprintf('Reinitialization:             %g',reinit));

    net=newff(minmax(inputTrainNormTrans),[numOfNeurons 1],...
{'tansig' 'purelin'},'trainlm');
    [net,trRec]=train(net,inputTrainNormTrans,targetTrain,...
[],[],val);

    predictedFromTest = sim(net,inputTestNormTrans);
    [fSlope,fIntercept,fR] = postreg(predictedFromTest,targetTest);

    fBias = sum(targetTest-predictedFromTest)./...
length(targetTest-predictedFromTest);
    fSEP = std(targetTest-predictedFromTest);
    fRPD = std(targetTest)/fSEP;

    if (trRec.vperf(end)<valMse)
       valMse=trRec.vperf(end)
       finalNet.net=net;
       finalNet.testData.inputTestNormTrans=inputTestNormTrans;
       finalNet.testData.targetTest=targetTest;
       finalNet.testData.transMat=transMat;
       finalNet.minFracOfVar=minFracOfVar;
       finalNet.explVar=(100-100*minFracOfVar);
       finalNet.numOfPCs=PCs;
       finalNet.numOfHLNeurons=numOfNeurons;
       finalNet.r=fR;
       finalNet.rSq=fR^2;
       finalNet.SEP=fSEP;
       finalNet.slope=fSlope;
       finalNet.intercept=fIntercept;
       finalNet.bias=fBias;
       finalNet.RPD=fRPD;
    end;
   end;
>>
>> % Display results.
>> finalNet
```

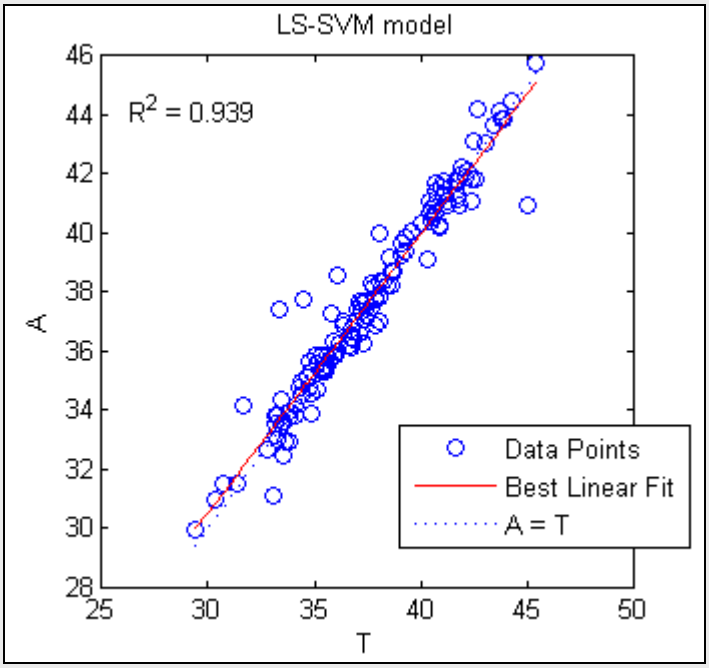
```
finalNet =

               net: [1x1 network]
          testData: [1x1 struct]
      minFracOfVar: 0.0025
           explVar: 99.7500
          numOfPCs: 23
    numOfHLNeurons: 3
                 r: 0.9656
               rSq: 0.9324
               SEP: 0.8997
             slope: 0.9246
         intercept: 2.8719
              bias: -0.0371
               RPD: 3.8454

>> close all;
>> pred=sim(finalNet.net,finalNet.testData.inputTestNormTrans);
>> figure, postreg(pred,finalNet.testData.targetTest);
>> title('ANN model');
>>
```



9. 9) Calibration using LS-SVM regression

Task: Develop and validate LS-SVM calibration model using data set "SbFossInfratecProteinCalData.mat".

```
>> % Load data file.
>> load SbFossInfratecProteinCalData.mat;
>>
>> % Preprocess spectra (2nd derivative).
>> spectra=savgol(spectra,5,3,2);
>>
>> % Divide data into calibration and validation sets.
>> % Use 75% of the samples for calibration and the other 25% for
>> % testing.
>>
>> inputTrain=spectra([1:4:end,2:4:end,4:4:end],:);
>> targetTrain=refData([1:4:end,2:4:end,4:4:end],:);
>>
>> inputTest=spectra(3:4:end,:);
>> targetTest=refData(3:4:end,:);
>>
>> % Normalize (autoscale) inputs and targets.
>> [inputTrainNorm,meanInputTrainNorm,...
stdInputTrainNorm]=auto(inputTrain);
>> inputTestNorm=scale(inputTest,meanInputTrainNorm,...
stdInputTrainNorm);
>> [targetTrainNorm,meanTargetTrainNorm,...
stdTargetTrainNorm]=auto(targetTrain);
>>
>> % LS-SVM regression.
>> gamSig2Range=[100 10; 1000000 100000]; % range for the two
>> % optimization parameters;
>> lssvmModel=initlssvm(inputTrainNorm,targetTrain,...
'f',1,0.1,'RBF_kernel','original');
>> lssvmModel=tunelssvm(lssvmModel,gamSig2Range,...
'gridsearch',{},'crossvalidate',...
{inputTrainNorm,targetTrain,5,'mse','mean','original'});
>> lssvmModel=trainlssvm(lssvmModel);
>> lssvmPred=simlssvm(lssvmModel,inputTestNorm);
>>
>> figure;
>> [lssvmSlope,lssvmIntercept,lssvmR]=postreg(lssvmPred',...
targetTest');
>> title('LS-SVM model');
>>
```



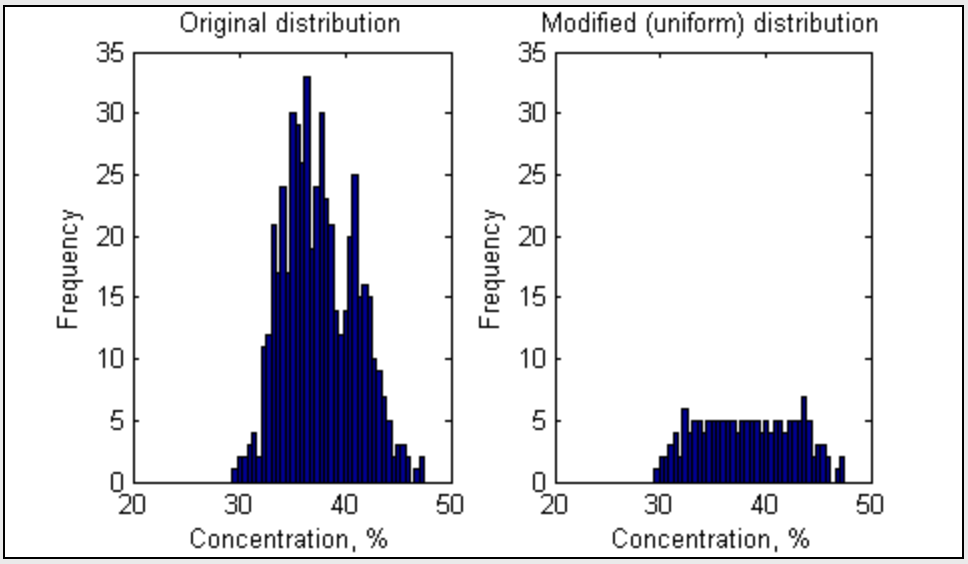

```
>> lssvmBias=sum(targetTest-lssvmPred)./length(targetTest-...
lssvmPred);
>> lssvmSEP=std(targetTest-lssvmPred);
>> lssvmRPD=std(targetTest)/lssvmSEP;
>> LSSVM.model=lssvmModel;
>> LSSVM.residuals=targetTest-lssvmPred;
>> LSSVM.r=lssvmR;
>> LSSVM.rSq=lssvmR^2;
>> LSSVM.SEP=lssvmSEP;
>> LSSVM.slope=lssvmSlope;
>> LSSVM.intercept=lssvmIntercept;
>> LSSVM.bias=lssvmBias;
>> LSSVM.RPD=lssvmRPD;
>> LSSVM.testData.pred=lssvmPred;
>> LSSVM.testData.actual=targetTest;
>>
>> % Display results.
>> LSSVM

LSSVM =

        model: [1x1 struct]
    residuals: [131x1 double]
            r: 0.9688
          rSq: 0.9386
          SEP: 0.8572
        slope: 0.9445
    intercept: 2.1860
         bias: -0.0976
          RPD: 4.0360
     testData: [1x1 struct]
>>
```

9. 10) Creating uniformly distributed data sets

Task: Obtain a uniformly distributed in reference values data set from normally distributed data in “SbFossInfratecProteinCalData.mat”.

```
>> % Load data file.
>> load SbFossInfratecProteinCalData.mat;
>>
>> % Assemble initial data matrix.
>> % 1st column is sample IDs, 2nd col. is reference values,
>> % the rest of the columns is spectral data.
>> data=[sampleIDs,refData,spectra];
>> % Sort rows by reference values.
>> data=sortrows(data,2);
>>
>> % Plot original distr. & define size of bin.
>> [rows,cols]=size(data);
>> binSize=floor(sqrt(rows/30));
>> num=10*binSize; % number of bins
>> conc=data(:,2);
>> subplot(1,2,1); hist(conc,40);
>> yAxisRange=get(gca,'YLim');
>> title('Original distribution');
>> xlabel('Concentration, %'); ylabel('Frequency');
>> minc=min(conc);
>> maxc=max(conc);
>> bin=(maxc-minc)/num;
>>
>> % Initialize bins.
>> for j=1:num
      bins(j).samps=[];
  end;
>>
>> % Assign samples to bins.
  for i=1:length(conc)
      b=ceil((conc(i,:)-minc)/bin);
      if b==0
         b=1;
      end;
      bins(b).samps=[bins(b).samps; data(i,:)];
  end;
>>
>> % Resample bins.
>> clear j;
>> for j=1:num
      numSamps=size(bins(j).samps);
      numSamps=numSamps(1);
      if numSamps>binSize
          step=floor(numSamps/binSize);
          bins(j).samps=bins(j).samps(1:step:end,:);
      end;
  end;
>>
```

```
>> % New uniform data set.
>> dataU=[];
>> clear j;
>> for j=1:num
      dataU=[dataU;bins(j).samps];
  end;
>>
>> % Plot new distribution.
>> subplot(1,2,2); hist(dataU(:,2),40);
>> set(gca,'YLim',yAxisRange);
>> title('Modified (uniform) distribution');
>> xlabel('Concentration, %');
>> ylabel('Frequency');
>>
```




```
>>
>> % dataU is a new data set uniformly distributed in reference
>> % values; 1st column of dataU is sample IDs, 2nd col. is
>> % reference values, the rest of the columns is spectral data.
>>
```

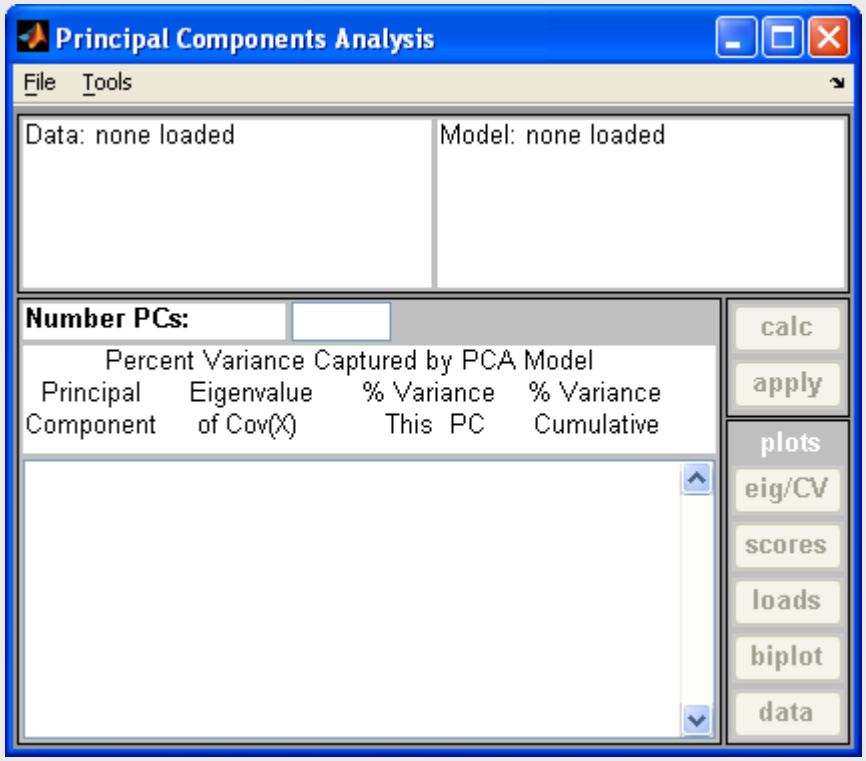
9. 11) Removal of spectral outliers

Task: Using principal component analysis, identify and remove spectral outliers from the data set in "SbFossInfratecProteinCalData.mat".

Solution:

a) In MATLAB command window:

```
>> % Load data file.
>> load SbFossInfratecProteinCalData.mat;
>>
>> % Start Principal Component Analysis tool.
>> pca
>>
```



b) Load **x**-data (spectral data) into PCA tool: File/Load Data, select variable spectra, and click Load.

c) Perform principal component analysis by clicking calc button.

d) Select number of PCs that explain 100% of variation (5 PCs) and click apply. The result of this operation is shown in Figure 68.

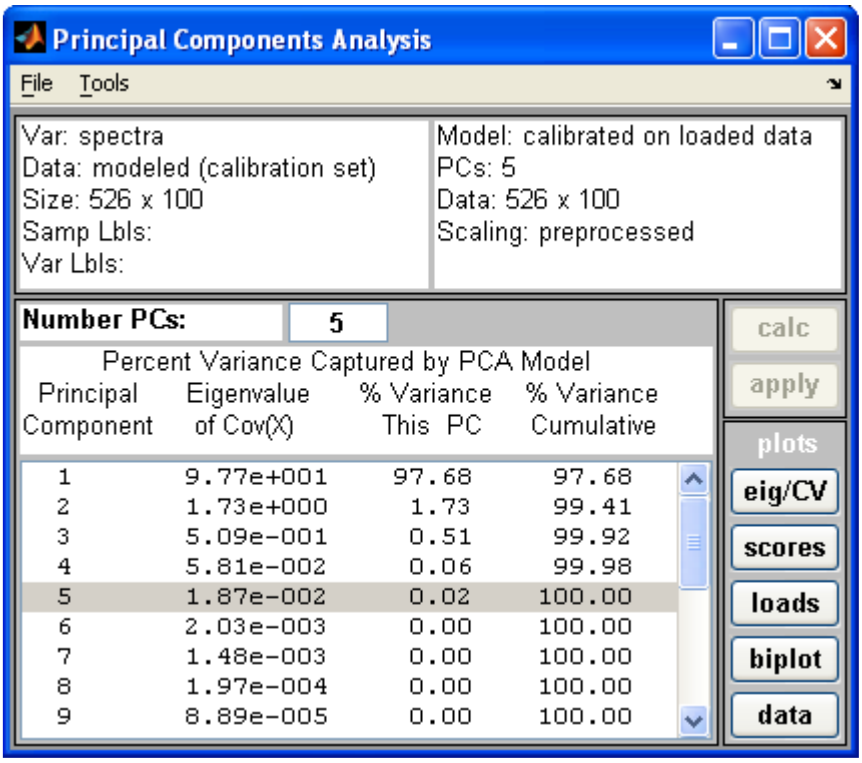


Figure 68. Selecting number of PCs.

e) Click scores button. Two additional windows will appear: plot of scores and Plot Controls dialog window.

f) In Plot Controls window, select Q Residuals for X-axis of the graph and Hotelling T^2 for Y-axis (Figure 69).

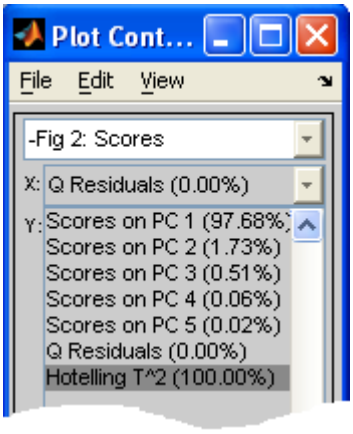


Figure 69. Selecting X- and Y-axes for the plot of scores.

g) In Plot Controls window, checkmark a box next to Conf. Limit 95%. Two dashed lines identifying 95% confidence limit intervals for Q Residuals and Hotelling T^2 will appear on the "plot of scores" (Figure 70).

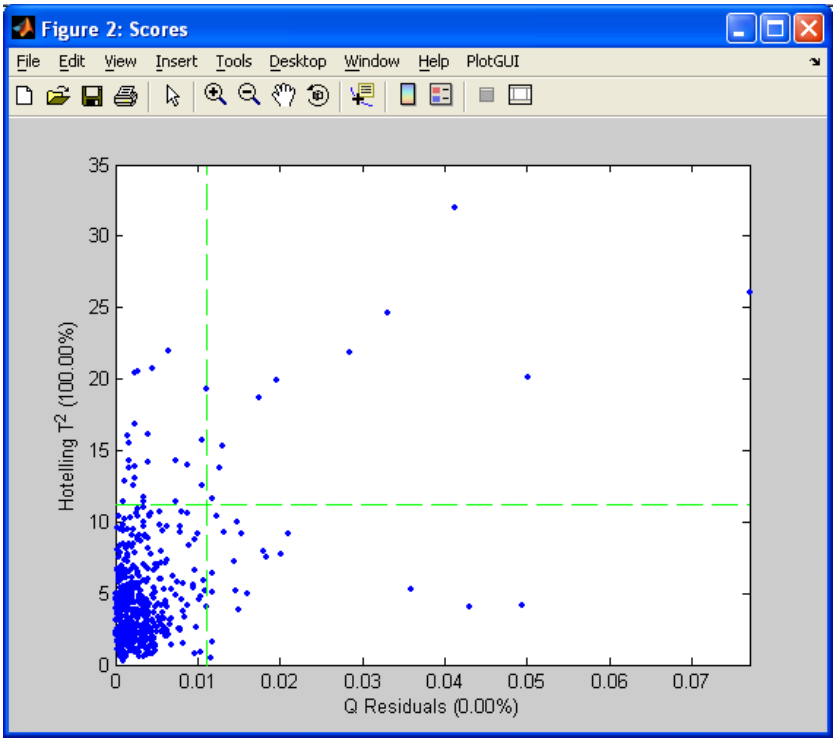


Figure 70. Plot of Hotelling T^2 vs. Q Residuals with 95% confidence limit intervals.

h) Using polygon selection tool (in Plot Controls, go to Edit/Selection Mode/Polygon and then click Select button), select data points beyond 95% limit (Figure 71).

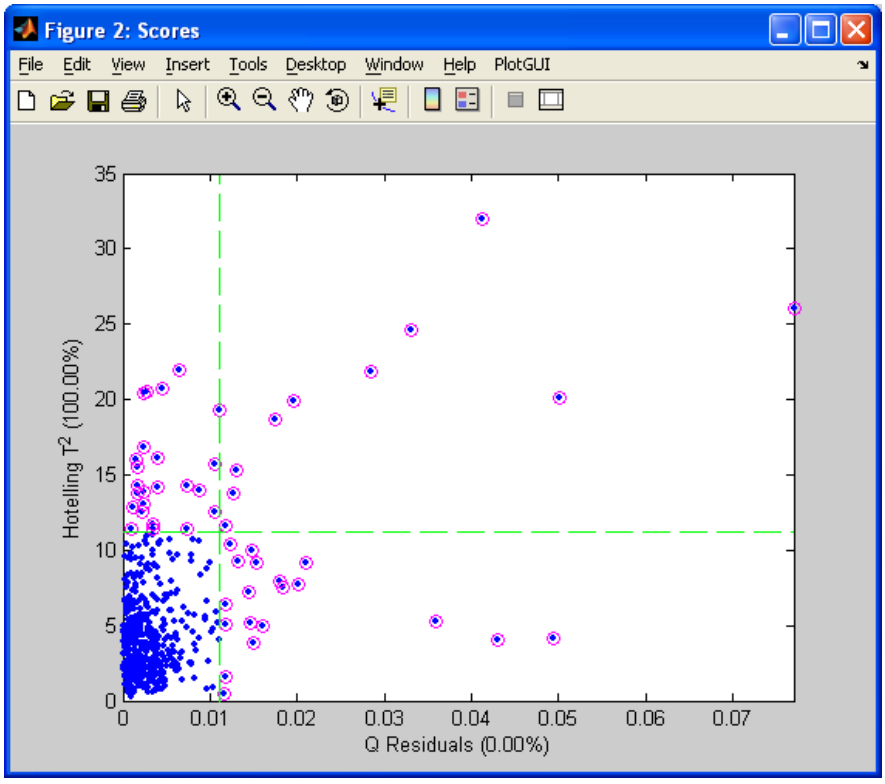


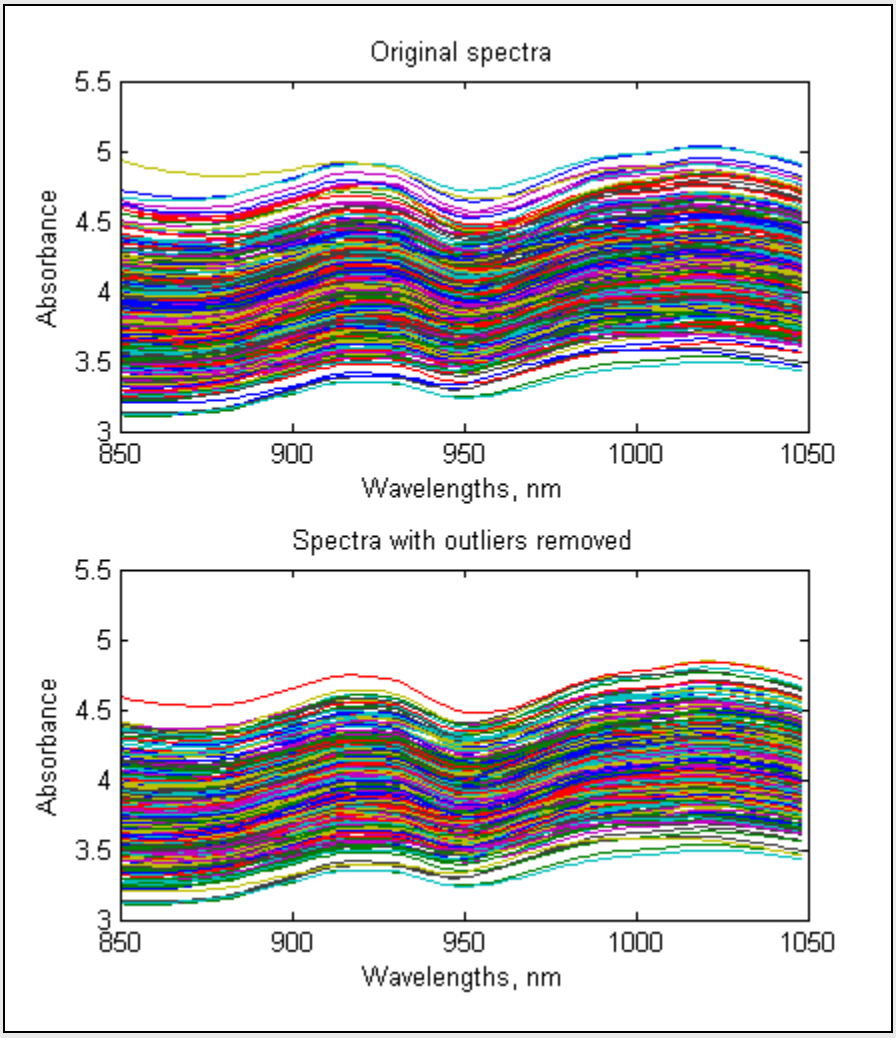
Figure 71. Plot of Hotelling T^2 vs. Q Residuals with data points beyond 95% confidence limit selected.

i) Exclude selected data points: in Plot Controls, go to Edit/Exclude Selection.

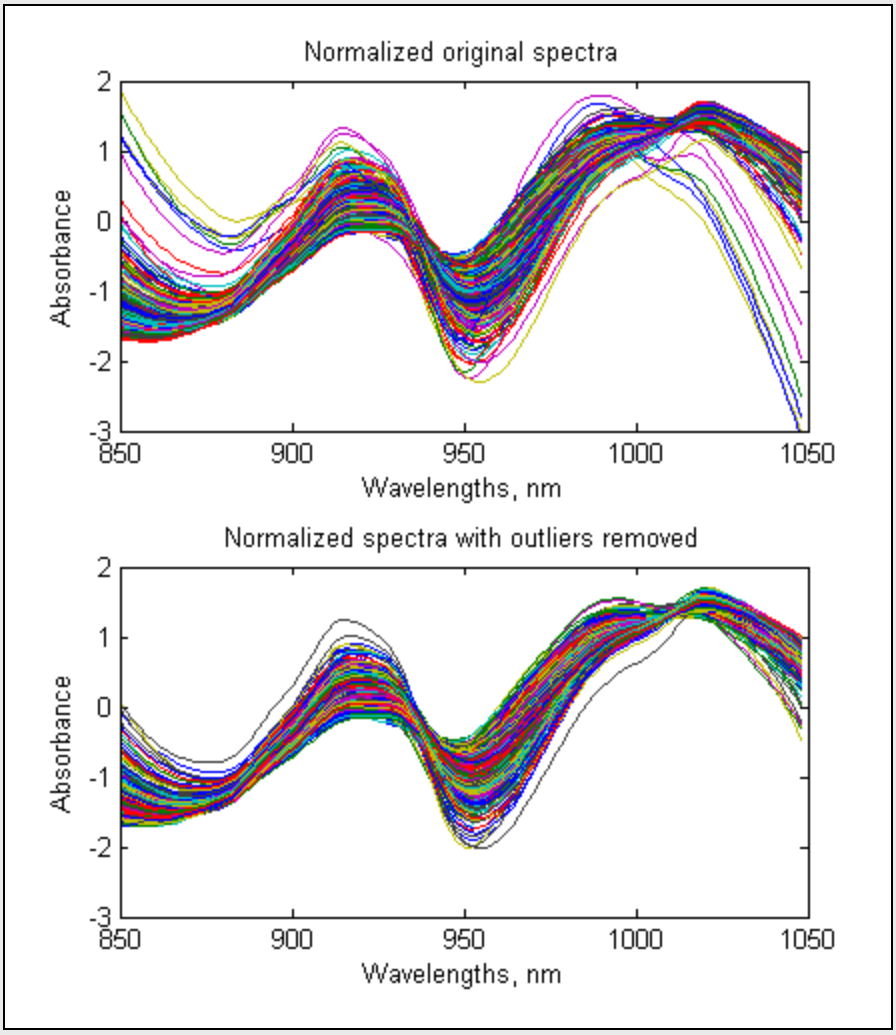
j) Save data to the MATLAB work space: File/Save Data, call new variable pcaData, and click Save.

k) In MATLAB command window:

```
>> includedSamps=pcaData.include{1};
>>
>> spectraNew=spectra(includedSamps,:);
>> refDataNew=refData(includedSamps,:);
>> sampleIDsNew=sampleIDs(includedSamps,:);
>>
>> figure, subplot(2,1,1),plot(wLens,spectra);
>> title('Original spectra');
>> xlabel('Wavelengths, nm');
>> ylabel('Absorbance');
>>
>> subplot(2,1,2),plot(wLens,spectraNew);
>> title('Spectra with outliers removed');
>> xlabel('Wavelengths, nm');
>> ylabel('Absorbance');
>>
```



```
>> % The difference between original and new data sets
>> % is more visible if the spectra are normalized using SNV.
>>
>> spectraSnv=snv(spectra);
>> spectraNewSnv=snv(spectraNew);
>> figure,subplot(2,1,1),plot(wLens,spectraSnv);
>> title('Normalized original spectra');
>> xlabel('Wavelengths, nm');
>> ylabel('Absorbance');
>> subplot(2,1,2),plot(wLens,spectraNewSnv);
>> title('Normalized spectra with outliers removed');
>> xlabel('Wavelengths, nm');
>> ylabel('Absorbance');
>>
```




```
>> % If a PLS calibration model is developed for a new data set,
>> % its validation r^2 improves from 0.928 (see example 8.7) to
>> % 0.936.
>>
```

10. RECOMMENDED READING

10.1) Books

A User-Friendly Guide to Multivariate Calibration and Classification by Tormod Næs, Tomas Isaksson, Tom Fearn, Tony Davies; NIR Publications, Chichester, UK, 2002, ISBN 0952866625.

Chemometric Techniques for Quantitative Analysis by Richard Kramer; Marcel Dekker, 1998, ISBN 0824701984.

Chemometrics: Statistics and Computer Application in Analytical Chemistry by Matthias Otto; John Wiley & Sons, 1999, ISBN 352729628X.

Data Fitting in the Chemical Sciences By the Method of Least Squares by Peter Gans; John Wiley & Sons, 1992, ISBN 0471934127.

Handbook of Near-Infrared Analysis by Donald A. Burns, Emil W. Ciurczak; Marcel Dekker, 2nd edition, 2001, ISBN 0824705343.

Multivariate Analysis of Quality: An Introduction by Harald Martens, Magni Martens; John Wiley & Sons, 2001, ISBN 0471974285.

Multivariate Calibration by Harald Martens, Tormod Naes, Tormod Ns; John Wiley & Sons, 1989, ASIN 0471909793.

Near-Infrared Applications in Biotechnology by Ramesh Raghavachari; Marcel Dekker, 2001, ISBN 0824700090.

Near-Infrared Spectroscopy: Principles, Instruments, Applications by H. W. Siesler, Y. Ozaki, S. Kawata, H. M. Heise; John Wiley & Sons, 2002, ISBN 3527301496.

Near-Infrared Technology in the Agricultural and Food Industries by Phil Williams, Karl Norris; American Association of Cereal Chemists, 2nd edition, 2001, ISBN 1891127241.

Neural Networks in Chemistry and Drug Design by Jure Zupan, Johann Gasteiger; John Wiley & Sons, 2nd edition, 1999, ISBN 3527297790.

Principles and Practice of Spectroscopic Calibration by Howard Mark; Wiley-Interscience, 1991, ISBN 0471546143.

Statistical Methods in Analytical Chemistry by Peter C. Meier, Richard E. Zünd; Wiley-Interscience, 2nd edition, 2000, ISBN 0471293636.

10. 2) Internet resources

Council for Near Infrared Spectroscopy:
http://www.idrc-chambersburg.org.

Glossary of NIR Terms:
http://www.asdi.com/ASD-600520_NIR-Glossary_Rev1.pdf.

NIR Publications:
http://www.nirpublications.com.

NIR Publications Discussion Forum:
http://www.nirpublications.com/discus.

Quantitative Analysis Using NIR and Chemometrics:
http://www.postech.ac.kr/class/chem441/exp8.htm.

Selection of a Multivariate Calibration Method:
http://minf.vub.ac.be/~fabi/calibration/multi/pages/start.html.

Spectroscopy Europe: The Tony Davies Column:
http://www.spectroscopyeurope.com/td_col.html.

Spectroscopy Magazine:
http://www.spectroscopymag.com/spectroscopy.

SpectroscopyNow:
http://www.spectroscopynow.com.

Theory and Principles of NIR Spectroscopy:
http://www.spectroscopyeurope.com/NIR_14_01.pdf.

11. REFERENCES

Aizerman, M., Braverman, E., Rozonoer, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25: 821–837.

Andrews, D.T., Wentzell ,P.D., 1997. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer. Analytica Chemica Acta, 350:341-352.

Armstrong, P. R., Maghirang, F. X. , Dowell, F. E., 2006. Comparison of dispersive and Fourier-transform NIR instruments for measuring grain and flour attributes. Applied Engineering in Agriculture, 22(3):453-457.

Axun technologies, 2005. Designing a miniature spectrometer. Technical note retrieved January 2010 from www.axsun.com/.../05-02-084a%20_designing_miniature_spectrometer%20Final.pdf

Corti, P. , Ceramelli, G. , Dreassi, E. , Mattiim, S., 1999. Near infrared transmittance analysis for the assay of solid pharmaceutical dosage forms. The Analyst, 124:755-758.

Bacci, M. , Bellucci, C. , Cucci, C. , Frosinini, C. , Picollo, M. , Porcinai, S. , Radicati, B. , 2005. Fiber optics reflectance spectroscopy in the entire VIS-IR Range: A powerful tool for the non-invasive characterization of paintings. In In: Vandiver, P. B. , Mass, J. L., Murray, A. (Eds.), Materials Issues in Art and Archaeology VII (vol. 852): Warrendale, PA.

Balas, C., 2009. Review of biomedical optical imaging – a powerful, non-invasive, non-ionizing technology for improving in-vivo diagnosis. Measurement Science technology, 20:1-12.

Berntsson, L-G. Danielsson, Folestad, S., 1998. Estimation of effective sample size when analyzing powders with diffuse reflectance near-infrared spectrometry. Analytica Chemica Acta, 364(1-3):243-251.

Bokobza, L., 1998. Near Infrared spectroscopy. Journal of Near Infrared Spectroscopy, 6 :3-17.

Börjesson, T. , Stenberg, B. , Schnürer, J., 2007. Near-Infrared spectroscopy for estimation of ergosterol content in barley: A comparison between reflectance and transmittance techniques. Cereal Chemistry, 84(3): 231-236.

Bouveresse E. Massart D.L., Dardenne P., 1994. Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: effects of different standardization samples, Analytica Chimica Acta, 297:405-416.

Bouveresse E., Hartmann C., Marrart L., 1996. Standardization of Near-Infrared Spectrometric Instruments. Analytical Chemistry, 68:982-990.

Brimmer, P. J., DeThomas, F. A., Hall, J. W., 2001. Method development and implementation of near-infrared spectroscopy in Industrial manufacturing processes. In: Williams, P. C., Norris, K., Near-infrared technology in the agricultural and food industries, AAC: St. Paul, MN.

Buchanan, R. R. , Honigs, D. E. , Lee, C. J. , Roth, W., 1988. Detection of Ethanol in Wines Using Optical-Fiber Measurements and Near-Infrared Analysis. Applied Spectroscopy, 42:1106-1111 .

Burns, D. A, Ciurczak, E. W. (Eds.), Handbook of Near-Infrared Analysis, 2nd ed., Marcel Dekkker, Inc., New York, NY, 2001.

Chung, H., Choi, S., Choo, J., and Lee, Y., 2004. Investigation of partial least squares (PLS) calibration performance based on different resolutions of Near Infrared Spectra. Bulletin of Korean Chemistry Society, 25(5):647-651.

Chang S.Y., Baughman E.H., McIntosh B.C., 2001. Implementation of Locally Weighter Regression to Maintain Calibrations of FT-NIT Analyzers for Industrial Processes, Applied Spectroscopy, 55:1199-1206.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association, 83:596-610.

Coats, D. B. , 2002. Is near infrared spectroscopy only as good as the laboratory reference values? An empirical approach. Spectroscopy Europe, 14(4):24-26.

Cogdill R.P., Hurburgh C.R., 2002. Local Chemometrics, 11th International Diffuse Reflectance Conference, Chambersburg, PA, USA.

Cogdill, R.P., Dardenne, P., 2004. Least-Squares Support Vector Machines for Chemometrics: An Introduction and Evaluation. Journal of Near Infrared Spectroscopy, 12:93-100.

Cogdill, R. P., Anderson, C. A., Delgado-Lopez, M. , Molseed, D. , Bolton, R. , Herkert, T., Afnan, A. M. , and Drennen III, J. K. , 2007. Process analytical technology case study

part I: Feasibility studies for quantitative near-infrared method development. AAPS PharmSciTech, 6(2):E262-E272

Davies, A.M.C., Grant, A., 1987. Review: near infrared analysis of food. International Journal of Food Science Technology 22: 191-207.

Davies , A.M.C., Britcher, H.V., Franklin, J.G., Ring, S.M., Grant, A., McClure, W.F., 1988. The Application of Fourier Transformed NIR Spectra to Quantitative Analysis by Comparison of Similar Indices (CARNAC). Mikrochimica Acta (Wien) 94:61-64.

Davies, A. M. C. , 2000. Honouring Herschel's discobery. Journal of Near Infrared Spectroscopy, 8(2):73-86.

Davies, A. M. C., 2005. An introduction to near infrared spectroscopy. Nir News 16(7): 9-21.

Davies, A. M. C., Fearn, T., 2006. Back to basics: calibration statistics. Spectroscopy Europe, 18(2):31-32.

Davies, A.M.C., Fearn ,T., 2006b. Quantitative analysis via near infrared databases: comparison analysis using restructured near infrared and constituent data-deux (CARNAC-D). Journal of Near Infrared Spectroscopy, 14:403-411.

Dean, T., Isaksson, T., 1993. Standardization: What is it and how is it done? Part 2. NIR News, 4(2):14-15.

Delwiche, S. R. , 1995. Single wheat kernel analysis by near-infrared transmittance: protein content. Cereal Chemistry, 72(1):11-16.

Despagne, F., Massart, L., 1998. Neural networks in multivariate calibration. Analyst, 123:157R-178R.

Despagne, F., Walczak, B., Massart, D.L., 1998. Transfer of calibrations of near-infrared spectra by neural networks. Applied Spectroscopy, 52:732-745.

Dixon, W. J., 1950. Analysis of extreme values. Annals of Mathematic Statistics, 21:488–506.

Domanchin, J. , Gilchrist, J. R. , 2001. Size and spectrum, Photonics, July 2001:12-118.

Drucker, H., Burges, C. J.C. , Kaufman, L. , Smola, A. , Vapnik, V., 1997. Support vector regression machines. Advances in Neural Information Processing Systems 9:155-161.

Dryden, G. Mc. , 2003. Near Infrared reflectance spectroscopy: Applications in deer nutrition. Publication No W03/007, Project No UQ 109A. Retrieved February 2009 from http://www.rirdc.gov.au/reports/DEE/w03-007.pdf

Duponchel, L., Ruckebusch, C., Huvenne ,J.P., Legrand, P., 1999. Standardization of near-infrared spectrometers using artificial neural networks. Journal of Near Infrared Spectroscopy, 7:155-166.

Fearn, T., 2001. Local or global?. NIR news, 12(3):10-11.

Fearn,T., 2002. Assessing calibrations: SEP, RPD, RER, and R2. NIR News 13(6):12-14.

Fearn, T., 2005. Chemometrics: an enabling tool for NIR. NIR News, 16(7):17-19.

Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D., Ferre, J., 2002. Transfer of multivariate calibration models: a review. Chemometrics and Intelligent Laboratory Systems, 64:181-192.

Fischer, D., Pigorsch, E., 2000. New developments in process control by spectroscopic methods in the polymer and plastics industry – near infrared miniature spectrometer and high temperature and pressure near infrared and raman probes. Paper presented at the third annual UNESCO school IUPAC conference on Macromolecules and materials science, Matieland, South Africa. Retrieved January 1010 from academic.sun.ac.za/unesco/PolymerED2000/Conf2000/Fischer.pdf

Garini, Y., Young, I. T., McNamara, G., 2006. Spectral imaging: Principles and applications. Cytometry, 69A(8) (2006):735 – 747.

Greensill, C. V., Walsh, K. B., 2000. Optimization of instrument precision and wavelength resolution for the performance of NIR calibrations of sucrose in water – cellulose matrix. Applied Spectroscopy, 54(3):426-430.

Grubbs, F.E. , 1950. Sample criteria for testing outlying observations. Annals of Mathematic Statistics, 21: 27–58.

Haaland, D. M., Thomas, E. V., 1988. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Analytical chemistry, 60:1193-1202.

Hammateenejad, B. , Akhind, M., Samar, F., 2007. A comparative study between PCR and PLS in simultaneous spectrophotometric determination of diphenylamine, aniline, and phenol: Effect of wavelength selection. Spectrochimica Acta, part A: Molecular and Biomolecular Spectroscopy, 67(3-4):958-965.

Herschel, F. W., 1800. Investigation of the powers of the prismatic colours to heat and illuminate objects. Philosophical Transactions of the Royal Society of London, 90: 255-329.

Holler, S. , Pan, Y. , Chang, R. K. , Bottiger, J. R., Hill, S. C. , Hillis, D. B. , 1998. Two-Dimensional Angular Optical Scattering for the Characterization of Airborne Microparticles. Optics Letters, 23:1489–1491.

Hopkins, D. W. , 2008. Using data pretreatments effectively, seminar at International Diffuse Reflectance Conference, Chamberburgh, PA.

Hubert, M., Rousseeuw, P. J., Van Aelst, S. , 2008. High-breakdown robust multivariate methods. Statistical Science, 23(1):92-119.

Hurburgh Jr. C. R., Rippke, G. R., 2008. Calibration, standardization and validation economics. Oral presentation at the International Diffuse Reflectance Conference (IDRC), Chamberburg, PA.

Ivanciuc, O., 2007. Applications of support vector machines in chemistry. Reviews in computational chemistry, 23: 291-400.

Kalivas, J. H., Gemperline, P. J., 2006.Calibration. In: Gemperline, P. J. (Ed.), Practical guide to chemometrics,CRC, Taylor and Francis group: Boca Raton,FL.

Kays, S. E., Shimizu, N. , Barton II , F. E., Ohtsubo, K. , 2005. Near-Infrared transmission and reflectance spectroscopy for the determination of dietary fiber in barley cultivars. Crop Science, 45:2307-2311.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics, 11:137-148.

Kennedy J., Eberhart R., 1995. Particle swarm optimization In: Proceedings of IEEE International Conference On Neural Networks, Perth, Australia, 1942–1948.

Kennedy J., Eberhart R.C., 2001. Swarm Intelligence, Academic Press, San Diego, CA, USA.

Kovalenko, I. V., Rippke, G. R. , Hurburgh, C. R. , 2006. Determination of amino acid composition of soybeans (Glycine max) by near-infrared spectroscopy. Journal of Agricultural and Food Chemistry 54:3485–3491

Lang, M., 1999. Diodes storm the tunable laser ranks, Photonics Tech Briefs. Retrieved January 2010 from http://www.ptbmagazine.com/articles/diodes0199/

Leardi R., Boggia R., Terrile M., 1992. Genetic Algorithms as a Strategy for Feature Selection, Journal of Chemometrics, 6:267-282.

Lipp, E. D. , 1992. Near-infrared spectroscopy of silicon-containing materials. Applied spectroscopy Reviews, 27(4):385-408.

Lu, J.X., Shen, Q., Jiang, J.H., Shen, G.L., Yu, R.Q., 2004. QSAR analysis of cyvlooxygenase inhibitor using particlw swarm optimization and multiple linear regression. Journal of Pharmaceutical and Biomedical Analysis, 35:679-687.

Luke, B.T., 1994. Evolutionary programming applied to the development of quantitative structure-activity relationships and quantitative structure-property relationships. Journal of Chemical Information and Modeling, 34:1279–1287.

Malinen, J., Känsäkoski, M., Rikola, R., Eddison, C. G., 1998. LED-based NIR spectrometer module for hand-held and process analyzer applications. Sensors and actuators B: Chemical, 51(1-3): 220-226.

Mark, H., Workman, J., 1988. A new Approach to generating transferable calibrations for quantitative near-infrared spectroscopy. Spectroscopy, 3(11):28-36.

Martens, H. , Næs, T., 2001. Multivariate Calibration by Data Compression. In: Williams, P., Norris, K. (Eds.), Near-Infrared Technology in the Agricultural and Food Industries, 2nd ed., AACC Inc.: St. Paul, MN.

Mcarthur, L., Greensill, C., 2007. Impact of resolution on NIR PLS calibration of kaolinite content with weipa bauxite. Measurement Science Technology, 18:1343-1347.

McClure, W. F., Moody, D. , Standfield, D. L. , Kinoshita, O. , 2002. Hand-held NIR spectrometry. Part II: An economical no-moving parts spectrometer for measuring chlorophyll and moisture. Applied Spectroscopy, 56(6):720-724.

Muñiz, R., Perez, M. A. , De la Torre, C. , Carlos, C. E., Corral, N. , Baro, J. A., 2009. Comparison of principal component regression (PCR) and partial least square (PLS) methods in prediction of raw milk composition by VIS-NIR spectrometry. Application to development of on-line sensors for fat, protein and lactose contents. Oral presentation proceedings for XIX IMEKO world congress of applied Metrology, Lisbon, Portugal, 229. Retrieved January 2010 from http://www.imeko2009.it.pt/Papers/FP_229.pdf

Naes, T. , Irgens, C. , Martens, H., 1986. Comparison of linear statistical methods for calibration of NIR instruments. Journal of the Royal Statistical Society Series C (Applied Statistics), 35(2):195-206.

Naes,T., 1987. The design of calibration in near reflectance analysis by clustering. Journal of Chemometrics, 1:121-134.

Næs, T., Isaksson, T., Kowalski, B., 1990. Locally weighted regression and scatter correction for Near-Infrared reflectance data. Analytical Chemistry, 62(7):664 - 673.

Naes, T., Isaksson, T., Fearn, T., Davies, T. (Eds.), 2002a. Outlier detection. In: A User-Friendly Guide to Multivariate Calibration and Classification, NIR publications: Chichester, UK.

Naes, T., Isaksson, T., Fearn, T., Davies, T. (Eds.), 2002b. Non-linearity problems in calibration. In: A User-Friendly Guide to Multivariate Calibration and Classification, NIR publications: Chichester, UK.

Naes, T., Isaksson, T. , Fearn, T. , Davies, T. (Eds.), 2002c. Selection of samples for calibration. In: A User-Friendly Guide to Multivariate Calibration and Classification, NIR publications: Chichester, UK.

Naes, T.,. Isaksson, T., Fearn, T., Davies, T. (Eds.), 2002d. Validation. In: A User-Friendly Guide to Multivariate Calibration and Classification, NIR publications: Chichester, UK.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. Applied Spectroscopy, 54(3):413-419

Norris, K. H., Hart, J. R. ,1965. Direct photospectrometric determination of moisture of grain and seeds.” Proceedings of 1963 International Symposium Humidity Moisture, vol.4 (pp. 19 – 25). Reinhold, NY.

Norris,K., Williams, P. C., 1984. 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 Chemistry, 61: 158-165.

Orman, B. A. , Schuman Jr., R. A., 1991. Comparison of near-infrared spectroscopy calibration methods for the prediction of protein, oil, and starch in maize grain. Journal of agricultural and food chemistry,39(5): 883–886.

Paynter, L.N., Hurburgh, C.R., 1983. Reference methods for corn moisture determination. American Society of Agricultural Engineers, 83:37-43.

Pou Saboya, N., 2002. Analisis de control de preparados farmaceuticos mediante espectroscopia en el infrarojo proximo. Phd dissertation (Barcelona: Universitat de Barcelona (UAB)).

Roger, J.M., Chauchard, F., Bellon-Maurel, V., 2003. EPO-PLS external parameter orthogonalization of PLS. Application to temperature-independent measurement of sugar content of intact fruits, Chemometrics and Intelligent Laboratory Systems, 66:191-204.

Sarraguca, M. C., Paulo, A. , Alves, M. M. , Dias, A. M. A. , Lopes, J. A. , Ferreira, E. C. , 2009. Quantitative monitoring of an activated sludge reactor using on-line UV-visible and near-infrared spectroscopy. Analytical and Bioanalytical Chemistry, 395(4):1159-1166.

Schumann, A. W., Meyer, J. H., 2000. Progress with the implementation of diode array near-infrared spectrometer for direct at-line analysis of sugarcane samples. Proceedings of South Africa Sugar Technology Association, 74:122- 123.

Shen, Q., Jiang, J.H., Jiao, C.X., Shen, G.L., Yu, R.Q., 2004. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. European Journal of Pharmaceutical Sciences, 22:145-152.

Shenk, J.S., Westerhaus, M.O., Templeton, W.C., 1985. Calibration transfer between Near Infrared reflectance spectrometers. Crop Science, 25:159-161.

Shenk , J.S., Westerhaus, M.O., 1993. Comments on Standardization: Part 2. NIR news, 4(5):13-15.

Shenk, J.S., Westerhaus ,M.O., Berzaghi, P., 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. Journal of Near Infrared Spectroscopy, 5:223-232.

Shi, B., Ji, B., Zhu, D., Tu, Z., Qing, Z., 2008. Study on genetic algorithms-based NIR wavelength selection for determination of soluble solids content in fuji apples. Journal of Food Quality, 31(2):232-249.

Short, S. M., Cogdill, R. P. , Anderson, C. A. , 2008. Figures of merit comparison of reflectance and transmittance near-infrared methods for the prediction of constituent concentrations in pharmaceutical compacts. Journal of Pharmaceutical Innovation, 3(1):41-50.

Siesler, H. W., Ozaki, Y., Kawata, S., Heise, H. M. (Eds), Near-Infrared Spectroscopy: Principles, Instruments, Applications.WILEY-VCH Verlag GmbH, Weinheim, Germany, 2002.

Siska J., Hurburgh C.R., and Siska P., 2001. The standardization of near-infrared instruments using master selection and Wiener filter methods. Journal of Near Infrared Spectroscopy, 9:97-105.

Smith, J. P., 2000. Product review: Spectrometers get small. Miniature spectrometers rival benchtop instruments. Analytical Chemistry, 72(19):653A-658A

Smola, A.J., Scholkopf, B., 1998. A Tutorial on Support Sector Regression. NeuroCOLT2 Technical Report Series, NC-TR-98-030, Royal Holloway College, University of London, UK.

Stark, E., Luchter, K., 2005. NIR instrumentation technology. NIR news, 16(7): 13-16.

Suykens, J.A.K., T. Van Gestel, J. De Brabanter, B. De Moor, Vandewalle, J., 2002. Least Squares Support Vector Machines, World Scientific, Singapore 2002.

Swayze, G. A., Clark, R. N., Goetz, A. F. H., Chrien, T. G., and Gorelick, N. S., 2003. The effects of spectrometer bandpass, sampling, and singnal-to-noise ratio of spectral identification using the tetracorder algorithm. Journal of geophysical research, 208(E9), 5105-

Swierenga, H., de Groot, P.J., de Weijer, A.P., Derksen, M.W.J., Buydens, L.M.C., 1998. Improvement of PLS model transferability by robust wavelength selection, Chemometrics and Intelligent Laboratory Systems, 41:237-248.

Tamburini, E., Vaccari, E. G., Tosi, S., Trilli, A., 2003. Near-infrared spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fiber probe. Applied Spectroscopy, 57(2):132-138.

Tarumi, T., Amerov, A. K., Arnold, M. A. , Small, G. W., 2009. Design considerations for Near-Infrared filter photometry: Effects of noise sources and selectivity. Applied Spectroscopy, 63(6):700-708.

Teppola ,P., Mujunen ,S.P., Minkkinen, P., 1999. Kalman filter for updating the coefficients of regression models. A case study from an activated sludge waste-water treatment plant. Chemometrics and Intelligent Laboratory Systems, 45:341-384.

Thermo Fisher Scientific, 2006. Advantages of Fourier-Transform Near-Infrared Spectroscopy. Application note 50771 retrieved January 2010 from: http://www.thermo.com/eThermo/CMA/PDFs/Articles/articlesFile_1195.pdf

Trygg, J., Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics, 16:119-128.

Vandegiste, B. G. M. , Massart, D. L., Buydens, L.M.C. , De Jong, S. , Lewi, P.J. , and Smeyers-Verbeke, J., 1998. Artificial neural Networks. In: Vandeginste, B. G. M., Rutan, S. C. (Eds.), Handbook of Chemometrics and Qualimetrics: Part B, Elsevier: Amesterdam, Holland.

Vapnik, V., Lerner, A., 1963. Pattern recognition using generalized portrait method. Automation and Remote Control, 24: 774–780.

Vapnik, V., Golowich, S. , Smola, A., 1997. Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing. In: Mozer, M., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9, MIT Press:Cambridge, MA.

Walczak, B., Massart, D. L., 1998. Multiple outlier detection revisited. Chemometrics and Intelligent Laboratory Systems, 41(1):1-15.

Wang, Y., Veltkamp, D.J., Kowalski ,R., 1991. Multivariate Instrument Standardization. Analytical Chemistry, 63:2750-2756.

Wang, W. , Paliwal, J. , 2006. Design and evaluation of a visible-to-near-infrared electronic slitless spectrograph. Science Technology, 17: 2698-2704.

Wash, K. B., Guthire, J. A. , Burney, J. W. , 2000. Application of commercially available, low-cost, miniaturized NIR spectrometers to the assessment of the sugar content of intact fruit, Australian Journal of Plant Physiology, 27(12):1175-1186.

Whetsel, K.B. ,1968. NearInfrared Spectrophotometry. Applied Spectroscopy Reviews, 2: 167.

Williams, P. C. , Sovering, D. C., 1993. Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. Journal of Near Infrared Spectroscopy, 1:25-32.

Williams, P. C. , 2001. Implementation of Near Infrared technology, in Near-Infrared technology in the agricultural and food industries. In: Williams, P. C., Norris, K. (Eds.), Near-infrared technology in the agricultural and food industries, AAC: St. Paul, MN.

Williams, P., Norris, K. (Eds.), Near-Infrared Technology in the Agricultural and Food Industries, 2nd ed., AACC Inc., St. Paul, MN, 2001.

Wise, B., no date. Properties of partial least squares regression and differences between algorithm. Brochure retrieved January 2010 from: http://www.eigenvector.com/Docs/Wise_pls_properties.pdf

Wold, S., 1975. Soft modeling by latent variables; the non-linear iterative partial least squares approach. In: Gani, J. (Ed.), Perspectives in probability and statistics, papers in honour of M.S. Barlett, Academic Press: London, UK.

Wold, S., Antti, H., Lindgren, F., Ohman ,J., 1998. Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems, 44:175-185.

Workman,J. J., 2005. An introduction to Near Infrared spectroscopy. Retrieved January 2010 from http://www.spectroscopynow.com/coi/cda/detail.cda?id=1881&type=EducationFeature&chId=2&page=1

Xie, Y., Hopke, P.K., Paatero , P., 1999. Calibration transfer as a data reconstrunction problem. Analytica Chemica Acta, 384:193-205.

Xing, J., Guyer, D., 2008. Comparison of transmittance and reflectance to detect insect infestation in Montmorency tart cherry. Computers and Electronics in Agriculture, 64(2):194-201.

Zeaiter, M., Roger, J.M., Bellon-Maurel ,V., 2006. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. Chemometrics and Intelligent Laboratory Systems, 80:227-235.

Zhu, Y., Fearn, T., Samuel, D., Dhar, A., Hameed, O., Bown, S.G., Lovat, L.B., 2008. Error removal by orthogonal subtraction (EROS): a customised pre-treatment for spectroscopic data. Journal of Chemometrics, 22:130-134.

Zomer, S., 2004. Classification with support vector machines. Homepage of chemometrics retrieved October 2009 from http://www.chemometrics.se/images/stories/pdf/nov2004.pdf