UNIVERSIDAD DE CÓRDOBA EN CONVENIO CON IOWA STATE UNIVERSITY

DEPARTAMENTO DE BROMATOLOGÍA Y TECNOLOGÍA DE LOS ALIMENTOS Y DEPARTAMENTO DE INGENIERÍA AGRÍCOLA Y BIOSISTEMAS

"ANÁLISIS MULTIVARIANTE PARA EL CONTROL DE CALIDAD EN MATERIAS PRIMAS DE USO AGROALIMENTARIO MEDIANTE ESPECTROSCOPÍA DE INFRARROJO CERCANO"

"Multivariate analysis for quality control of agrifood materials using near infrared spectroscopy"

Tesis Doctoral

Mariana Soto Cámara

Directores:

Rafael Moreno Rojas

Antonio J. Gaitán Jurado

Charles R. Hurburgh

TITULO: *ANÁLISIS MULTIVARIANTE PARA EL CONTROL DE CALIDAD EN MATERIAS PRIMAS DE USO AGROALIEMTNARIO MEDIANTE ESPECTROSCOPÍA DE INFRARROJO CERCANO.*

AUTOR: *MARIANA SOTO CÁMARA*

© Edita: Servicio de Publicaciones de la Universidad de Córdoba.
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

Departamento de Bromatología
y Tecnología de los Alimentos y
Departamento de ingeniería agrícola y biosistemas

"ANÁLISIS MULTIVARIANTE PARA EL CONTROL DE CALIDAD EN MATERIAS PRIMAS DE USO AGROALIMENTARIO MEDIANTE ESPECTROSCOPÍA DE INFRARROJO CERCANO"

TESIS

Para aspirar al grado de Doctora por la Universidad de Córdoba presentada por la Licenciada en Ciencias Ambientales

La doctoranda

Fdo: Mariana Soto Cámara

V°B° Los Directores

Fdo: Prof.Dr. Rafael Moreno Rojas	Fdo: Dr. Antonio J. Gaitán Jurado	Fdo: Prof. Dr. Charles R. Hurburgh Jr.

Charles R. Hurburgh

Digitally signed by Charles R. Hurburgh
DN: cn=Charles R. Hurburgh, o=Iowa State University, ou=Agricultural Engineering, email=tatry@iastate.edu, c=US
Date: 2013.04.09 09:50:08 -05'00'

2013

Departamento de Bromatología
y Tecnología de los Alimentos y
Departamento de ingeniería agrícola y biosistemas

ANTONIO GAITÁN JURADO, Dr. Ingeniero agrónomo, RAFAEL MORENO ROJAS director de departamento de Bromatología y Tecnología de los Alimentos de la Universidad de Córdoba y CHARLES R. HURBURGH JR. Profesor del departamento de ingeniería agrícola y biosistemas de la Universidad de Iowa.

INFORMAN:

Que la tesis titulada "ANÁLISIS MULTIVARIANTE PARA EL CONTROL DE CALIDAD EN MATERIAS PRIMAS DE USO AGROALIMENTARIO MEDIANTE ESPECTROSCOPÍA DE INFRARROJO CERCANO", de la que es autora Dña. Mariana Soto Cámara, realizada en el periodo de 2008-2013 bajo nuestra dirección, cumple las condiciones académicas exigidas por la Legislación vigente para optar al título de doctor por la Universidad de Córdoba.

Y para que conste a los efectos oportunos firman el presente informe en Córdoba a 5 de abril del 2013.

Fdo: Prof. Dr. Rafael Moreno Rojas	Fdo: Dr. Antonio J. Gaitán Jurado	Fdo: Prof. Dr. Charles R. Hurburgh Jr.

Charles R. Hurburgh

Digitally signed by Charles R. Hurburgh
DN: cn=Charles R. Hurburgh, o=Iowa State University, ou=Agricultural Engineering, email=tatry@iastate.edu, c=US
Date: 2013.04.09 10:02:39 -05'00'

2013

Acknowledgments

Es difícil agradecer en pocas palabras a las personas que han formado parte de mi vida durante este período tan importante. Esas personas que me han ayudado y animado para que siga hacia adelante. A todos, mi más sincero agradecimiento.

Todo esto no lo habría conseguido sin la ayuda de mi familia. Mis padres, que siempre han estado a mi lado apoyándome y aconsejándome, gracias por haber hecho de mi lo que soy. A mi hermano y Trini por su apoyo incondicional y Águeda que aún siendo más pequeña que yo ha sabido darme sabios consejos. A Pablo, por la paciencia que ha tenido conmigo y todo el cariño que me ha dado cuando más lo he necesitado. Os quiero a todos.

A las personas que me han dado la oportunidad de embarcarme en este viaje, Juan y Fide, mil gracias, os estaré eternamente agradecida.

A mis compañeros con los que he compartido cafés, vivencias y preocupaciones y que han pasado a ser una parte importante de mi vida: Lourdes, Curro, Elena, Miguel y Javi. A Juani por toda la ayuda que me ha aportado y a la que he mareado tantas veces.

A la gente que ha hecho que durante mis estancias fuera no echase tanto de menos mi casa: Tara y Jeff, Lidia, Susana, Luis, Aldane, Glen y Nancy. Y a los que me han dado la oportunidad de trabajar con ellos y han compartido su experiencia, a Vincent, Juan y a la Profesora Pilar Barreiro.

Por la suerte que tengo de cantidad y calidad amigos y que por desgracia no puedo mencionar a todos. Gracias por darme momentos inolvidables, cariño y ser pacientes conmigo.

A mi director Antonio por sus ánimos y por el tiempo dedicado. A Charlie por haber confiado en mí y preocuparse como un padre. A Rafael, por estar ahí cuando lo he requerido.

Por último hacer mención a la importante labor que efectúan los organismos INIA-IFAPA (Subprograma FPI-INIA), cofinanciado por FSE fondos (Programa Operativo FSE de Andalucía 2007-2013_"Andalucía se mueve con Europa"), que me ha permitido llevar a cabo esta investigación. Y a la Red Andaluza de Experimentación Agraria (RAEA) por los datos aportados.

A mis padres

Contents

Figures Index

Tables Index

Abbreviations

AACC: American association for clinical chemistry

ANNs: Artificial neural networks

AOAC: Association of analytical chemists

AOCS: American oil chemists' society

CC: Correct classification

CT: Samples that were run at cold temperature (5ºC)

E: Extensibility of the flour

FTIR: Fourier transform infrared

g: Grams

HPLC: High-performance liquid chromatography

ICC: International association for cereal science

IFAPA: Instituto de investigación y formación agraria y pesquera

IR: Infrared region of electromagnetic spectrum

ISO: International organization for standardization

LVQ: Learning vector quantization

Max: Maximum

MD: Mahalanobis distance

Min: Minimum

MIR: Mid infrared region of electromagnetic spectrum

ml: Millilitres

MPLS: Partial least squares modified

MSC: Multiplicative scatter correction

mt: Millions tonnes

NIRS: Near Infrared Spectroscopy

No TC: Calibration group without temperature compensation included

NTEP: National type evaluation program

O: Non treated plants

OSC: Orthogonal signal correction

PCA: Principal component analysis

PLS: Partial least squares

R: Reflectance

R^2: Coefficient of determination

r^2: Coefficient of determination of cross validation

RER: Ration error of Range

RMS: Root mean square

RPD: Residual predictive deviation

RT: Samples that were run at room temperature (22ºC)

SD: Standard deviation

SDS-PAGE: Sodium dodecyl sulphate polyacrylamide gel electrophoresis

SEC: Standard error of calibration

SECV: Standard error of cross validation

SEP: Standard error of prediction

SIMCA: Soft independent modelling of class analogies

SNV: Standard normal variate

SVM: Support vector machine

T: Tenacity of the flour

T: Transmittance

T: Treated plants

TC: Temperature compensation samples

UV: Ultraviolet region of electromagnetic spectrum

Vis: Visible region of electromagnetic spectrum

W: Baking strength

WMSC: Weighted multiplicative scatter correction

WT: Samples that were run at warm temperature (45ºC)

Summary

Seguridad y calidad alimentaria son uno de los conceptos más demandados actualmente en la industria agroalimentaria. La mayoría de análisis de control de los productos alimentarios se lleva a cabo mediante métodos tradicionales (vía húmeda). Los principales problemas relacionados con este tipo de análisis son el consumo de tiempo para la obtención de los resultados de una sola muestra, el coste del análisis, así como la limitación en cuanto a su implantación en la línea de producción o en el campo, entre otros.

Paralelamente al desarrollo e innovación tecnológica, numerosos métodos han sido implementados para la determinación, evaluación y control de la calidad de los productos agroalimentarios en las últimas décadas. Estos métodos están basados en la detección de varias propiedades tanto físicas como químicas correlacionadas con ciertos factores cualitativos de los productos. Uno de los métodos más difundido y aún en desarrollo debido a su gran aplicabilidad, es la espectroscopía de infrarrojo cercano (tecnología NIRS, Near Infrared Spectroscopy). Han pasado más de 20 años desde su primera introducción como potente herramienta hecha por Karl Norris en el análisis de la composición de los cereales.

El planteamiento de esta tesis nace de la necesidad, cada vez mayor, del control de los parámetros de calidad de los productos agroalimentarios de manera rápida y precisa. La categorización del trigo en función de su calidad o el valor añadido que adquiere la soja según el porcentaje de proteína o grasa presente en una determinada variedad ha llevado al estudio de la aplicación de la espectroscopía de infrarrojo cercano en dichos productos.

El objetivo general de la investigación ha consistido en la aplicación de la tecnología NIRS para la determinación de parámetros de calidad en muestras de

trigo y soja. Como consecuencia, este estudio ha dado lugar al desarrollo de cuatro trabajos:

- "Development of robust soybean NIR calibration models with high variability and temperature compensation in the base data". Enfocado al desarrollo de calibraciones robustas añadiendo variabilidad tanto instrumental como ambiental en el colectivo de muestras.

- "Adjusting NIR calibrations for intact soybean with different path length instruments". Dado la velocidad con que evoluciona esta instrumentación, existe la necesidad de implantar las calibraciones desarrolladas en modelos de instrumentación antiguos en nuevos más versátiles. En este capítulo se explican métodos de transferencia de calibraciones entre distintos instrumentos.

- "Application of Near Infrared Spectroscopy technology and hyperspectral NIR imaging for the detection of fungicide treatment on durum wheat samples". Este estudio muestra la capacidad de estas dos herramientas a la hora de discriminar entre muestras de trigo duro que han sido infectadas por agentes fitopatógenos, afectadas en su matriz y como consecuencia en la calidad del trigo.

- "Application of NIRS in authentication of bread wheat varieties from southern Spain". En este capítulo busca la utilidad de esta herramienta en la discriminación entre variedades de trigo harinero, factor importante en la industria harinera.

El potencial de la tecnología NIR para control de la calidad tanto en soja como en trigo junto con la aplicación de herramientas quimiométricas, queda patente en este trabajo. La aplicación de la espectroscopia de infrarrojo cercano puede ser usada para determinar, caracterizar y cuantificar parámetros de calidad de dichos productos.

Food safety and quality are currently the most popular concepts in the food industry. Usually, most control analyses of food products are carried out by conventional methods (wet chemistry). However, some of the main negative issues of these methods are: they are time consuming in order to obtain the results of a single sample, the raising price and the limitation on its implementation in the production line or in the field, among others.

At the same time to the technological innovation and development, during the last decades many methods have been implemented for the identification, assessment and quality control of food products. These methods are based on the detection of various physical and chemical properties correlated with certain product quality factors. One of the most widespread due to its wide applicability is the near-infrared spectroscopy (NIRS technology, Near Infrared Spectroscopy). It has been over 20 years since its first introduction as a powerful tool made by Karl Norris in the analysis of the composition of the grains.

The approach of this thesis arises from the increasing need of fast and accurate analyses of quality parameters control on food products. The categorization of wheat in terms of quality and the added value acquired by the percentage of soy protein or fat in a particular variety has led to the study of the application of near infrared spectroscopy in these products.

The general objective of the research has been the application of NIRS technology for the determination of quality parameters in wheat and soybean samples. As a result, this study has led to the development of four chapters:

- "Development of robust soybean NIR Calibration Models with temperature compensation and high variability in the data basis." This chapter was focused on the development of robust calibrations by adding in the group of samples instrumental and environmental variability.

- "Application of Near Infrared Spectroscopy and hyperspectral NIR imaging technology for the detection of fungicide treatment on durum wheat samples". This study shows the ability of both tools on discriminating between durum wheat samples that have been infected by pathogenic agents. Because of this the quality of wheat grains has been affected consequently by presenting modifications in the kernel matrix.

- "Application of NIRS in authentication of bread wheat varieties from southern Spain." This study seeks the usefulness of this tool in discriminating between bread wheat varieties, important factor in the flour industry.

- "Adjusting for intact soybean NIR calibrations with different path length instruments". Given the speed with which this instrumentation evolves. There is an obvious need of using calibration models developed in olden instrumentation in a new more versatile instrument. This work describes methods of transfering calibrations between instruments.

The potential of NIR technology together with chemometrics tools for quality control in both soybean and wheat is shown in this thesis. The application of near-infrared spectroscopy can be used to identify, characterize and quantify quality parameters such products.

Chapter I

Introduction

1.1. Agrarian Production

1.1.1. *General introduction to Agrarian Production. A brief history.*

The origins of agriculture are independently linked to several locations across the world, beginning about 12,000 years ago. Vavilov identified eight centers of domestication, based primarily on patterns of crop diversity, which Harlan reduced to three the main centers, which were relatively small areas, and another three rather diffuse regions [1]. These three "centers" correspond to the Near East (beginning about 9,500 BP); Meso-America (about 7,200 BP) and the Peruvian highlands (about 6,500 BP) (Figure I-1).

At the end of the last Ice Age, humans became more sedentary mainly due to environmental changes and to a restructured distribution of resources. New ecological niches offering potential resources arose, and new habitats were exploited. At this stage, a major advance occurred with the harvesting of wild grains, hand in hand with the technological advances required to roast and prepare cereals for human consumption [2].

The historical trajectory of agriculture was transformed from symbolic to social and eventually to economic domestication: it began to be seen as a long-term process, with importance in social and economic areas. Around 10,000 years ago, agriculture began to profoundly change how human communities worked, many hunter-gatherers moved over to farming and agricultural societies began to supply food instead of relying on hunting and gathering. These changes led to an increase in population density and an increased infant survival rate, but also to a greater spread of density-dependent diseases and other pathological conditions related to poor diets. What this means is that agriculture helped people to survive but it did not make them healthier (this was only achieved after the Industrial Revolution) [3].

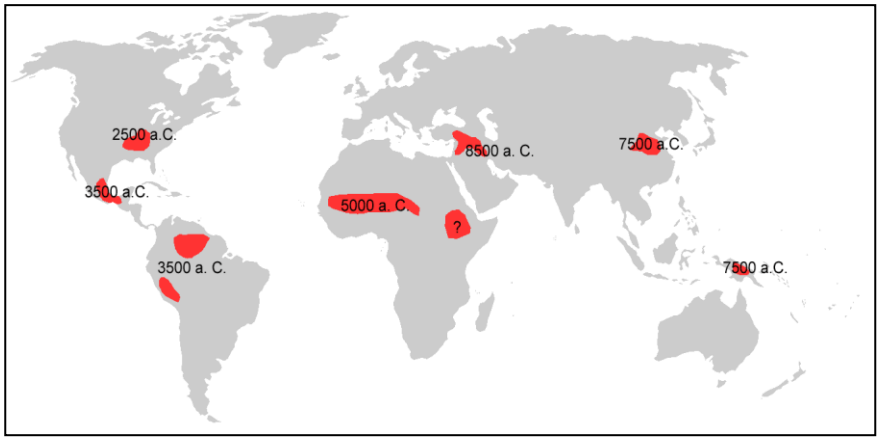


Figure I-1. Centers of origin of food production. (Source: Jared Diamon).

Throughout history, depending on their needs, humans have developed a variety of ways to acquire or produce foods, in order to feed

themselves and to enhance the health and growth of their populations. In this context, five main strategies have been recognized as types of food systems: hunting-gathering, horticultural, pastoral, intensive agriculture and industrialism.

For about 99% of their history, human survival was solely based on hunting and gathering. This means of survival is the oldest and most widely distributed, and it was not until 10,000 years ago that societies started to employ agriculture. The evolution from the hunter-gatherer lifestyle to food production allowed people to settle down instead of migrating to follow seasonal shifts in wild food supplies. The toolkit was quite simple, formed by light killing weapons, spears, atlatls, bows and arrows. For food collection, a digging stick and a slab of bark or a simple wooden bowl sufficed. This shift, known by archaeologists as "the broad spectrum revolution", led to the domestication of plants as the most important source of subsistence. Societies based on agriculture were characterized by the use of simple tools and the lack of plows and animal traction [4].

However, the main advance in food production was the domestication of animals. Tamed animals became a mainstay of human existence and development. The animals got used to human presence, became dependent on a human environment and were a useful tool for traction or transport and a vital source of milk, meat and hides.

The beginnings of farming were marked by the biological domestication of plants and animals. The previous role of hunter-gatherers was to kill animals; now they tried to ensure their survival. These technical and demographic changes led to much more complex societies, which eventually brought us to the industrial revolution. Farming implements and

tools were replaced by industrial-scale agricultural machinery: harvesters, combines, tractors, cultivators and milking machines. Organic fertilization systems gave way to chemical fertilizers. These two elements, machinery and chemical fertilizers, destroyed the traditional function of livestock, which turned into a mere transformer of food resources.

According to this view, the Industrial Revolution culminated an evolutionary process that had seen society develop from a traditional agricultural economy to one where mechanized production processes produced manufactured goods which were intended for sale on a large scale. In the period before the Industrial Revolution, food was produced exclusively for home consumption; after that, the objective was to sell it to distant regions and countries.

During the period of the Industrial Revolution, there was an enormous expansion in many areas. These changes have a particular bearing on food production, regulations, and on quality control too [5]. Although the principles governing food control can be traced back to times of the earliest societies, quality control and the basic uses of statistical principles are modern concepts [5], [6], [7].

Nowadays, quality food control, in a scientific way, evaluates certain factors (technological, physical, chemical, microbiological, nutritional and sensorial) and their properties (texture, color, taste, etc) to assess the wholesomeness of food. The purpose of controlling quality is to maintain acceptable standards and limits of tolerance according to demand and to reduce supply costs [5].

1.1.1.1. Wheat and Soybean crops

Up until 10,000 years ago, humans did not eat cereals. Around that time, whole grains became part of the human diet, and humans began to move from place to place to find seasonal seeds, a major food source, and eat them as they ripened. When the supply of seasonal seed was limited, they moved on to another food until the next harvest arrived. For the last 3,000-4,000 years, the majority of the world population has relied upon whole grains as the main proportion of its diet [8].

Of the total of 195,000 species of edible flowering plants, less than 0.1% or fewer than 300 species are used for food; and of these, only 17 make up 90% of the food supply [9]. Recently, cereal and legumes have become the main dietary component for the majority of the world population. They play the leading role in feeding some human populations, especially in underdeveloped countries, where they make up around 50-80% of their sustenance, while in developed countries they only represent around 15% [10]. Derpsh 2010 [11], reported that the world would have to increase its food production by around 70% by 2050 to meet the needs of its growing population. These changes would reflect, above all, the rising consumption of cereals in developing countries, whose average is likely to rise from 2,680 kcal per person in 1997-1999 to 2,850 kcal in 2015 and close to 3,000 kcal in 2030 [12].

Cereals are the most important source of the total food consumption (in terms of calories) and will continue to be the most essential part of the human diet by far. In developing countries, the per capita average food use is now 173 kg, providing 56% of total calories, compared with 141 kg and 61%

in the mid-1960s. The figure of around 173 kg has remained fairly constant since the mid-1980s [12].

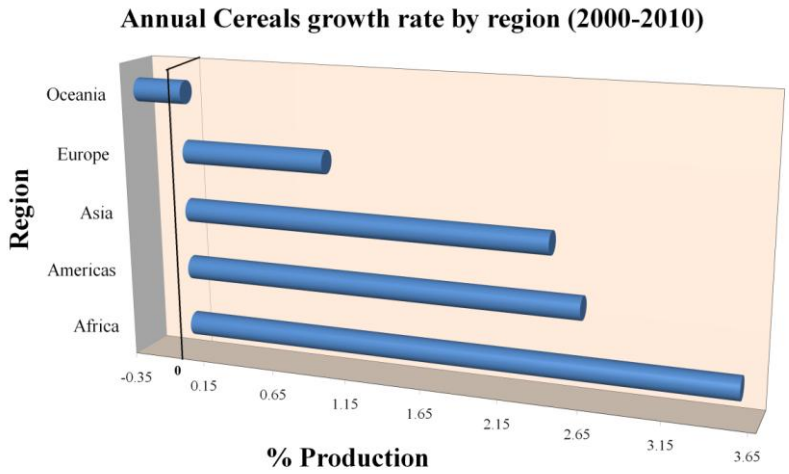


Figure I-2. Annual Growth Rate of the cereal production from the year 2000-2010. Source: FAOSTAT

Figure I-2 represents the percentage of annual growth rate of cereals all over the world from the crop seasons 2000 to 2010. There is a clear tendency to growth, especially in Africa; America and Asia behave in a similar way, and both have a growth rate close to 2.5%. Europe follows the same trend as America and Asia but with a slower rate: 1%. Oceania, however, shows negative growth. The demand for food grows continuously, the main reason being improved living standards and diets and the influence of technological improvements [13]. The major cereal grains include wheat, rice and maize, with wheat the most commonly consumed.

The wheat plant (*Triticum* ssp) covers more of the earth's crop surface than any other crop; after rice, it is the main food crop. The cultivation of this cereal began about 10,000 years ago from wild species harvested by ancient hunter-gatherers in South-east Asia. There are two main groups which make up most of the varieties: common wheat and durum wheat. Common or bread wheat (*Triticum aestivum)* accounts for some 95% of all the wheat consumed in the world today; the rest is made up of durum wheat (*T. turgidum* ssp. *Durum*), which is used in pasta and semolina products. The percentages of the principal uses of this crop are 70% used for food, 19% for animal feed and 11% for industrial applications. The wheat plant is an annual grass crop which can reach 1.2 meters in height. The stems are erect, cane-like in structure and hollow inside, except at the nodes. The growth of the shoot is apical and is produced by the stretching of the tissues above the knots (meristem); the leaves grow from these knots. Like all grasses, the wheat plant consists of two parts: the sheath that surrounds and protects the stalk or meristem growth zone and the limb which is elongated and has parallel ribs.

The flowers are not very showy, have no petals or sepals and end in spikes. Each spike consists of a main shaft or rachis where the spikelets are distributed laterally. These consist of a main shaft with filaments at the end enclosing flowering glumes which later begin to ripen. These flowering glumes are protected by two bracts: the inner is called the palea and the outer, the lemma. The latter is topped by a beard that gives the ear of wheat its feathery appearance (Figure I-3).

Figure I-3. Wheat ear. Source: Wikipedia

The grain is usually between 5 and 9 mm in length and can be of different shapes from nearly spherical to long, narrow and flattened. The wheat grain contains 2-3% germ, 13-17% bran and 80-85% mealy endosperm. Every part of the whole-wheat grain provides elements which are necessary for the human body. Starch and gluten provide heat and energy; phosphates and other mineral salts are present in the inner brand coats; fibre (which helps with bowel movement) is present in the indigestible portion; vitamins B and E are found in the germ; and proteins are present, which are necessary to help build and repair muscular tissue [14].

In terms of calories and proteins, wheat contributes to the world diet more than any other cereal crop. The average chemical composition of whole wheat grain consists of: 16% proteins, 2% fat, 68% carbohydrates, 11% dietary fibre and 3% minerals and other components.

In addition, oil crops have also played an important role in increasing food consumption in developing countries. According to future projections,

45% of additional calories in the period up to 2030 may come from these products [12]. This crop sector has been one of the most dynamic in world agriculture. Over the past 20 years, the sector has grown annually 4.3%, two points above the average for all agriculture foods [12]. 75% of the world's oil crop production is made up of four crops (oil-palm, soybeans, rapeseed and sunflower seed),

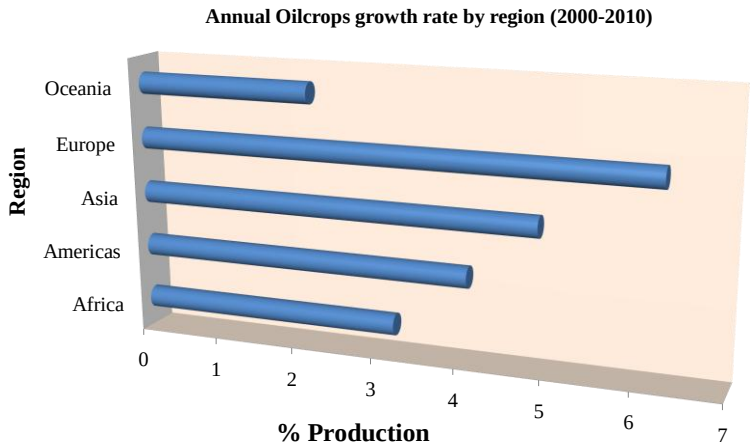


Figure I-4. Annual growth rate of cereal production in the period 2000-2010. Source: FAOSTAT

Figure I-4 represents the percentage of the annual growth rate in oil crops. Europe tops the list (with 6.26 %), 1.4 % ahead of Asia (4.89%) and followed by America (4.07%), Africa (3.19%) and Oceania (2.16%). There is great demand for oils with the potential for rapid expansion of production and those oil crops with high protein content are in high demand as animal feed.

Soybean (*Glycine max* L.) is one of the world's most economically important crops. Its origins are widely believed to lie in China, 4,000-5,000 years ago, and it was introduced in Europe around the year 1712 by a German botanist, Engelbert Kaempfer [15]. The composition of this seed makes the plant unique; it has excellent nutritive value, with fibre, a small portion of saturated fatty acids, 20% of oil content and ±40% protein content, which makes it suitable for a wide variety of applications, such as human consumption, livestock feed or industrial purposes. Around 90% of the world's soybean production is found in the United States of America, followed by Brazil and Argentina [12]. It is an annual spring-summer cycle crop, and soybean plants can reach between 50 cm and 1.80 m in height. The seed is produced in pods of 4 -6 cm of length which each contain 2 or 3 soybean grains (Figure I-5).

Figure I-5. Pods containing soybean grains. Source: Wikipedia

From a nutritional point of view, the Soybean plant has some advantages over other crops: it is N_2 fixative, needs low P inputs, can grow on low pH and high Al soils, and tolerates flooding [16]. It is an exceptional source of high quality, cholesterol-free protein, and low in saturated fat. Fibre, which helps to reduce the risk of bowel disease, is its second largest component.

1.1.1.2 The need for quality parameters in crops.

The growth in the production and consumption of wheat and soybeans has greatly increased the demand for quality in the original products and their derivations [17]. It is important to determine the chemical, biological and physical properties of the grains and to do this, new technologies are needed to reach the standards established by government regulations rapidly, accurately and safely.

For instance, traditional methods such as Kjeldahl (to determine protein levels in wheat); oven drying (to measure moisture in wheat); combustion (to measure oil in soybeans); ether extract (to measure oil in soybeans) etc, are considered tedious and time-consuming. There is a growing need for new technologies to reduce analytical time and provide reliable results.

NIRS (Near Infrared Spectroscopy) is a relatively new, promising form of technology. It meets the challenge of the increased demand for quality in the food industry, from the determination of parameters (such as protein, moisture, etc) to authentication and certification. Nowadays,

technology is developing rapidly and the daily search goes on for new applications to use in the agrofood industry.

The main body of this dissertation focuses on the use of this technology with certain crop seeds to help in the quality control of agricultural products.

Before starting with the applications of this technology for agricultural products, we will start with a brief introduction about the basic principles and chemometric aspects of this technology.

1.2 NIR Spectroscopy

1.2.1 Introduction and review

The Near Infrared (NIR) spectrum covers the range from 780-2500 nm. It is situated between the middle infrared (2500-50000 nm) and UV-Vis (280-780 nm) ranges (Figure I-6)

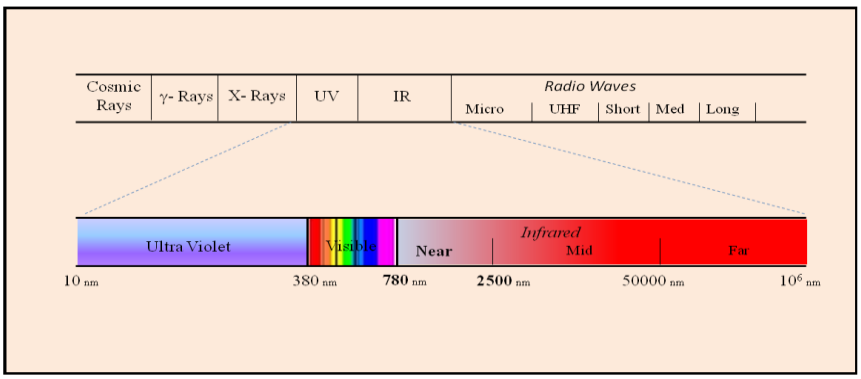


Figure I-6. Representation of the Electromagnetic Spectrum

Historically, the discovery of the NIR region is ascribed to Frederick William Herschel, who, in 1800, demonstrated the existence of radiation under the visible spectrum. He placed a glass prism in front of a slit cut in a window blind. Mercury-in-glass thermometers with blackened bulbs were positioned to measure the heat associated with different positions in the dispersed spectrum displayed on a horizontal surface (Figure I-7). Herschel realized that the temperature reached the maximum just beyond the red end of the spectrum. However, he recorded only the beginning of what we now call the NIR region - it was not until 1840 that the first measurements of NIR absorption bands were made by his son John Herschel. He put solar radiation through a glass prism and plotted the perceived evaporation of alcohol on a blackened sheet of paper as a detector [18].

Figure I-7. Herschel and his experiment to demonstrate the existence of IR radiation

The spectral characteristics of the NIR region delayed it use. In 1881 Abney and Festing [19] documented the first NIR spectra of organic liquid

over the 700-1200 nm range; Abney associated individual absorption bands in the NIR spectra with smaller functional groups within complex organic molecules [18]. However, it was not until midway through the twentieth century that this technology really came into its own. In 1954, Wilbur [20] showed the spectral representation of different organic products over the 700-3500 nm range.

In the 1960s, with the work of Karl Norris leading the United States Department of Agriculture (USDA), acceptation for NIR technology as an analytical method began to grow [21]. There was a growing need for rapid measurements and the quantitative determination of moisture, protein and oil [22].

This growth of work involving Near Infrared has occurred recently: until 1970, there were only about 50 papers written on work involving NIR, whereas now in the 21st century, the use of this technology has expanded considerably. One of the advantages of the NIR region is the vast range of materials that can be studied, often without special preparation [18].

Nowadays, NIR technology has an immense number of applications, including the pharmaceutical industry [23], [24], [25] and [26]; medicine [27], [28] and [29]; the textile industry [30], and the petrochemical industry [31].

In agricultural sciences, this technique has become an important tool for routine application. The measurement of parameters such as protein, oil, moisture, ash, crude fibre in grain [32], [33], [34], [35] and [36] or cotton [37], [38] are widely employed; as are measuring sugar content in grape [39], strawberry [40] and orange [41] or measuring the qualities of whole olives [42]. It is used not only for whole produce, but also for control processes of

animal feed and foodstuffs, such as checking the quality of potato crisps [43], ammonium monitoring [44], animal feed [45] and [46], checking eggs for freshness [47] and checking meat [48].

The development of new chemometric tools for data management and the reduction of the instrumental component, as well as the improvement in accuracy, have meant that this technology has perfectly suited current needs, not to mention the advances in NIR instrumentation, such as Hyperspectral NIR imaging, which allow us to derive significantly more information from each picture [49], [50] and [51], NIR-microscopy [52] and [53], or NIR portable [54] and [55].

1.2.2 Basics of Near Infrared Spectroscopy (NIRS)

Near Infrared Spectroscopy (NIRS) is based on the absorption of the electromagnetic radiation over the 780-2500 nm wavelength. The most important absorption bands in this region are related to overtones and combinations of fundamental vibrations in the Infrared (IR) region of the -CH, -NH, -OH functional groups [56] and [57]; most of the peaks observed in an NIR spectrum come from stretching movements of these bonds, appearing between 10 and 1000 times weaker than the main bands (Figure I-8) [58]. The origin of the absorption bands in the NIR is the same as with the MIR. IR radiation is absorbed by a molecule when the incident radiation coincides with the difference between two energetic states. After that, owing to vibrating movements, a change in the dipole moment has to occur. The dipole moment is the intensity of strength of attraction between two atoms. The spectroscopic phenomenon of energetic absorption by matter is caused by molecular rotations in the far infrared, fundamental vibration bands in the

middle infrared and, lastly, a combination of fundamental bands and overtones in the near infrared.

Two types of vibrations can occur in response to light energy hitting the bonds: stretching and bending vibrations. Stretching vibrations involve movements along the axes of the bonds. These types of vibrations tend to happen at higher frequencies (shorter wavelength) [59] and can be either symmetrical or asymmetrical. Bending vibrations can involve changes between the bond angles of atoms or groups of atoms and the rest of the molecule [60].

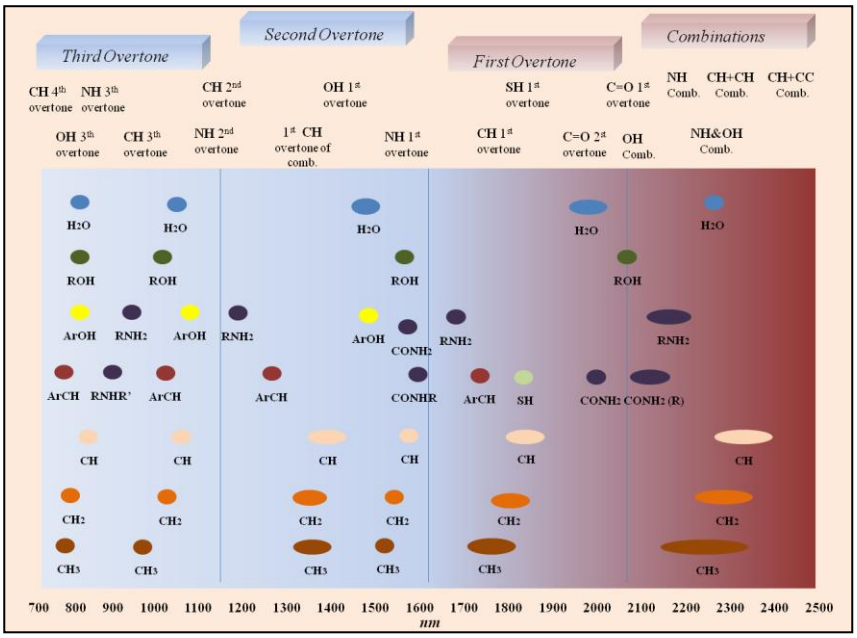


Figure I-8. Combination bands and overtones of the NIR region.

1.2.3 NIR measurements

One of the advantages of NIR spectroscopy is its capability of measuring samples of different types. The ideal mode for NIR measurement depends on the optical properties of the sample; changes in the particle size produce changes in the NIR spectrum.

Three ways (Figure I-9) of capturing the NIR spectrum, depending on sample presentation, can be considered: *reflectance, transmittance and transflectance.*

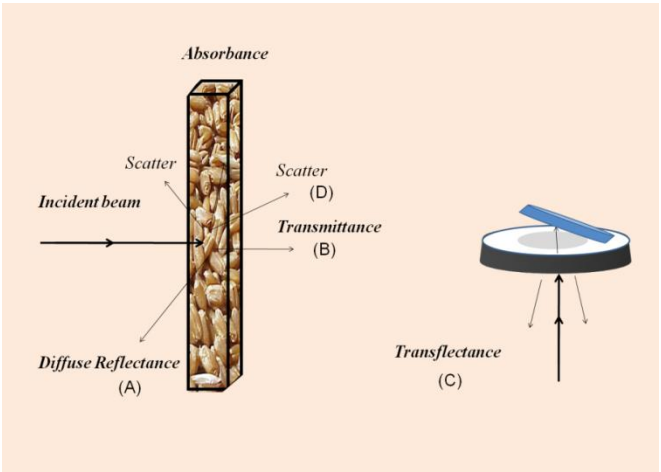


Figure I-9. Different spectral registration, depending on the sample-light interaction.

Diffuse Reflectance (A):

Diffuse reflectance is the sum of absorption and disperse radiation. This type of measurement is mainly used for thick, solid samples. The radiation impacts on the surface of the sample and is returned to the detector,

which is situated in front of the sample at an angle which minimizes the scatter radiation. The signal is expressed as:

$$A_{ap} = \frac{\log 1}{R} = ac$$

Where R is the relative reflectance, ($R=R_{sample}/R_{standard}$) is a proportionality constant and c a concentration [61].

Transmittance (B):

Transmittance is used for both liquid and solid samples. The incident radiation passes through the sample and follows the Beer-Lambert law, the fundamental law of quantitative absorption spectroscopy. This law is based on the direct proportionality for a given wavelength between absorbance and concentration of the chemical constituent [62] and its equation is:

$$T = \frac{I}{I_0} = \varepsilon l c$$

Where transmittance is represented by T, I_0 is the incident energy and I the intensity of energy emerging from the sample. ε is the characteristic absorptivity of a specific component at a specific wavenumber, l is the optical

path length and c the concentration of the light-absorbing chemical species in the sample [63].

Transflectance (C):

This means of registering is a combination of reflectance and transmittance modes, and is useful for thick, liquid samples. Transflectance measures the radiation as it passes through the sample, collides with a reflective surface and passes through the sample again to the detector.

Scatter (D):

This process is produced by the redistribution in all directions of the energy incident on the particles of matter.

1.2.4 NIR instrumentation

The need for a rapid, versatile response in analysis has made NIR a suitable technique for the assessment of a wide variability of samples. At a basic level, all NIR instrumentation is made up of:

Radiation source.

NIR energy is provided by a source. The most commonly utilized are tungsten filaments and halogen lamps, although other sources of radiation are available, such as LEDs (Light Emitting Diodes) [64]. Ration power is an

important factor when selecting NIR instruments, since it influences the depth of penetration of the sample.

Wavelength selection device.

This allows us to collect the spectral information required for the different wavelengths, splitting the polychromatic light into discrete wavelengths. This system is the main cause of variability in NIR instrumentation.

Depending on the device used, two types of NIR instruments can be found on the market:

Dispersive (such as grating monochromators, or diode array detectors), which use gratings or prisms to achieve wavelength selection, and are divided into pre-dispersive and post-dispersive instruments.

With pre-dispersive instruments, monochromatic light is sent through optical fibre to gratings. The dispersed light is transmitted through the sample, after which, light passes through a slit and reaches the detector. With post-dispersive instruments, light is sent directly to the sample. The light returning from the optical fibre is directed to a grating where it is dispersed and passes through a slit placed before detector.

Non–dispersive instruments, these do not use any gratings or prisms: isolation of a spectral band is achieved without wavelength dispersion by using optical absorption, fluorescence, reflection or scattering. It is also achieved by the use of an interference filter based on multiple beam interference.

Fourier Transform Infrared (FTIR) spectrometers use a Michelson interferometer to modulate the intensity of the infrared radiation as a function of frequency, and then employ Fourier transform algorithms to convert the resulting time-dependent spectrum into a standard wavenumber.

Detectors.

Detectors receive the radiant energy from the sample and transform it into an electrical signal. There are various types of detectors used in NIR instrumentation, depending on the spectral range: Silicon (Si) (which covers most of the visible range and a limited part of NIR), Lead Sulphide (PbS), Indium Gallium Arsenide (InGaAs), Indium Arsenide (InAs), Lead Selenide (PbSe) or Indium Antimonide (InSb).

1.3 Chemometrics

1.3.1 Definition

The evolution of NIR instrumentation generates a great deal of information in a short period of time, all of which needs to be processed. The conversion of this information into useful data requires mathematical and statistical tools, and the discipline that governs how these tools design and extract the information is called Chemometrics.

One of the many definitions of this discipline was given by Vandeginste (1988) [65], who postulated that "*Chemometrics is a chemical discipline that uses mathematics, statistics and formal logic to design or select optimal experimental procedures; to provide the maximum relevant*

chemical information by analyzing chemical data; and to obtain knowledge about chemical systems".

The spectral signal obtained during NIR analysis is formed by complex matrices containing relevant information about the composition of the sample, which makes the use of chemometric techniques such as multivariate analysis essential for analyzing and managing this immense data set. Chemometrics is a practical tool that performs several functions during spectral management in NIR analysis, such as improving the signal quality, applying multivariate techniques for clustering samples based on their qualities or using multivariate techniques that search for a quantitative relationship between the analytical signal and any property of the sample. The most common multivariate analyses can be explained as follows:

1.3.2 Signal Pretreatments

Once the NIR spectra are obtained, they can be influenced by changes in the spectrum not related to the component being studied which produce interference in the signal (which normally appears as baseline shift, tendency or sometimes a curvature). Because of this, it may first be necessary to apply a spectral pre-treatment in order to improve the signal/noise ratio in the component in question.

Frequent sources of spectral variation are chemical compound interactions; scattering produced by solid or liquid particles in suspension, spectral noise produced by instrumentation and overlapping of the spectra coming from overtones and combinations of the spectral bands [66] and [67].

Spectral pre-processing consists of a set of mathematical steps performed on spectral data before developing a calibration model. Some of the spectral pre-treatment used in our studies are explained below.

1.3.2.1 *Spectral Smoothing*

The aim of this method is to reduce mathematically random noise and improve the signal/noise ratio [68]. Smoothing is very useful to eliminate the noise generated by fast changes in signal amplitude. The most widespread methods of smoothing are the Savitzky-Golay method and the Fourier transform.

1.3.2.*2 Mean centering*

This consists of calculating the average spectrum of all the spectra in the sample set and then subtracting the result from each spectrum. Mean centering is a common pre-treatment before performing a calibration model or principal component analysis.

1.3.2.3 *Derivatives*

This is one of the best ways of removing baseline effects and enlarges the differences between band widths and overlapping spectra. The first and second derivative are the most frequently used in NIR technology. The first derivative is a very effective method for removing baseline offsets, and the second corresponds with wavelength movements, such as changes in the

slope of the curve. Normally in NIR technology, the latter is more commonly used, which allows us to differentiate overlapping peaks and to reduce lineal differences.

1.3.2.4 Multiplicative Scatter Correction (MSC)

This procedure corrects the shift produced by scattering with the linearization, obtained by minimum least squares, of each spectrum to an original spectrum [69]. The key to this method is that all spectra must be corrected using the same original spectrum (also called the "ideal" spectrum). MSC calculates the average spectrum from all the data in the sample set and uses it as the "ideal" spectrum.

The spectral responses in each spectrum are used to calculate a linear regression against the corresponding point in the "ideal" spectrum. The resultant corrected MSC spectrum has the slope and offset values previously subtracted, and conserves the chemical information while the differences between spectra are minimized.

This pre-treatment works well with chemically-similar spectra, but a sample composition with a wide range of variability would not give the desired results.

1.3.2.5 Standard Normal Variate (SNV)

SNV has the same base as MSC, the correction of the multiplicative and additive effects, with the difference that a SNV operates on a per-spectrum basis. With this pre-treatment, no "ideal" spectrum is required; the

scattering is removed by normalizing each spectrum by the standard deviation of the responses across the entire spectral range. The result obtained is a transformed spectrum with mean 0 and standard deviation 1, and the shift correction is along the vertical axis [70].

1.3.2.6 Orthogonal Signal Correction (OSC)

OSC is based on the fact that the majority of the spectral variance in a NIR data set is of little or no analytical value. It is therefore a specific pre-treatment for a selective collective of samples and also for the analyte. Variance that is orthogonal to the property under study is removed from the data set.

1.3.3 Multivariate analysis techniques

One of the advantages of NIR is its ability to develop quantitative and qualitative analytical methods for a wide variety of parameters that require minimal or non-processing analysis.

After applying spectral pre-treatment in order to minimize the effects produced by the sample, or by instrumental or environmental factors, the process of qualitative (PCA) or quantitative models (calibration methods) follows (Figure I-10).

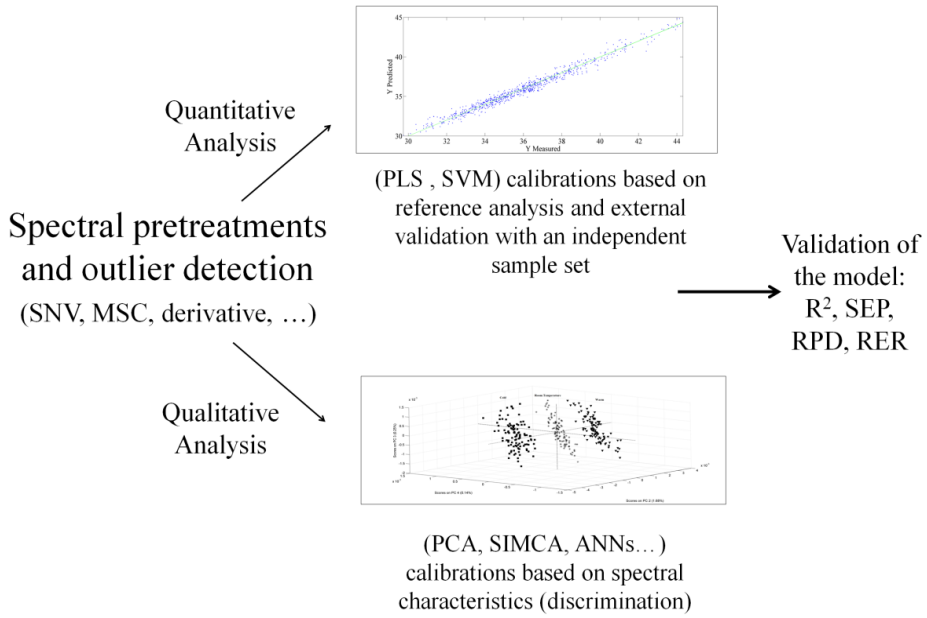


Figure I-10. Procedure of qualitative and quantitative analysis of NIR

1.3.3.1 Qualitative analysis

1.3.3.1.1 Principal Component Analysis (PCA)

The PCA method consists of a reduction of the spectral data into a new dimensional space. It is one of the most widely used methods of variables reduction, allowing us to obtain the maximum knowledge without losing any relevant information. The reduction of these variables reveals the existence of similarities or differences between groups or a group of samples.

The new space is defined by axes called Principal Components (PCs) or Eigenvectors. These PCs are orthogonal. The projection of the samples onto the PCs called scores (Figure I-11), and the variables are called loadings.

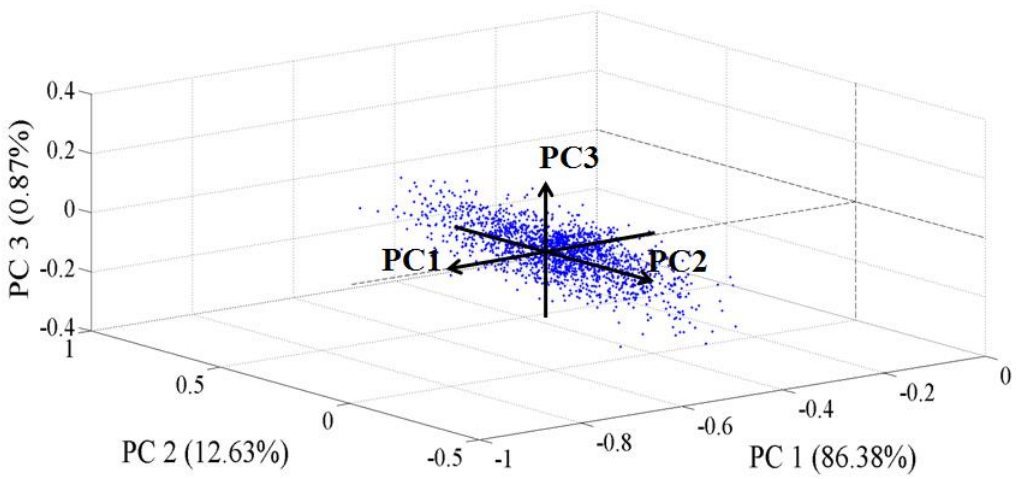


Figure I-11. 3D example of principal component analysis

The main principle is to approximate the main data matrix (X) to a product of two matrices of smaller dimensions, given by scores (T) and loadings (P):

$$X_{nxp} = T_{nxd} \times L_{dxp} + E$$

Where:

- X_{nxp} is the original matrix with n objects (samples) and p variables (absorbance values),
- T_{nxd} is the matrix for the new coordinates of each point according to the new axis (PCs)
- L_{dxp} is the matrix that relates the PCs with the absorbance values,
- E is the matrix for the residuals.

During the PCA, it is vital to select the optimal number of PCs, in order to divide the component which generates noise from the main spectroscopic information. This type of analysis is extremely useful, particularly in the first steps of the process, to have a general idea of the data when performing an exploratory analysis.

1.3.3.1.2 Soft Independent Modelling of Class Analogies (SIMCA)

SIMCA uses the modelling properties of the PCA technique, with each class being processed independently. For each class, a PCA is performed and new samples are tested against each of the models with a cross validation performance, to check which it most closely resembles. The resultant analyses correspond very closely to each group defined by the principal component.

SIMCA is sensitive to the quality of the data used to generate the principal component model [71].

1.3.3.1.3 Artificial Neural Networks (ANNs)

The design of ANNs was inspired by the structure of a real brain, although the processing elements and the architecture used now have

developed far beyond their biological inspiration [72]. It tries to reproduce the simplest connection system existing between the neurons of a human brain.

It consists of many interconnected neurons in the network, each of which is able to receive information signals, process them and send them back with an output signal.

1.3.3.1.4 Support Vector Machine (SVM)

The SVM is based on the classification performance by constructing an N-dimensional hyperplane that optimally separates the data into two or more categories. The aim of this technique is to find the optimal hyperplane that separates clusters of one feature (vectors) in one plane from others in another part of the plane. The vectors near the hyperplane are known as support vectors. SVM analysis looks for the line or curve that maximizes the margin between the categories [69].

1.3.3.2 Quantitative Analysis

This analysis is associated to the relationship between the concentration (Y) and the reflectance or transmittance measurements (X).

1.3.3.2.1 Partial Least Squares (PLS)

In NIRS analysis, large quantities of variables are given in response to each sample; unfortunately these variables cannot be assigned to only one analyte. The calibration process allows us to establish a relation between the instrumental response and the desired property.

PLS is probably the most frequently used regression method [73], and it allows us to calibrate the desired component without the necessity of knowing the rest of variation sources.

The PLS process is bilinear (like PCA), and the matrix formed by the X and Y data can be constructed from linear combinations of their scores and loadings [59]. The main objective of PLS is based on the calculation of the reduction in variables, where the X (spectral) and Y (analyte) matrices are decomposed simultaneously by:

$$X = TP^T + E$$

$$Y = UQ^T + F$$

Where:

- T is the matrix of the factor scores for the new coordinates of the data points in the X space,
- P^T is the matrix that contains the X loadings vectors,
- E is the matrix for X residuals,
- U is the matrix for the factor scores for the new coordinates of the data points in Y,
- Q^T is the matrix that contains loadings from the factorization of Y, and

  F is the matrix for Y residuals.

The search for the maximum correlation between spectra and the analyte is the main characteristic of this simultaneous decomposition [74].

1.3.4 Statistical regression analysis of the results

Once a calibration model is performed, it must be tested in order to know if the model has a suitable prediction capability. The statistics allow us to determine the fitness of a particular calibration and give some indication of the ability of the model to predict the analyte in question.

Statistic analysis establishes the relation between the real and predicted ($\bar{Y}$) values of Y. The regression line obtained by these $\bar{Y}$ values versus the real Y values is given by the equation:

$$\bar{Y} = a_1 Y + a_0$$

Where, a_1 and a_0 are the slope and bias, respectively. Slope and bias mark the differences between the calibration sample set and the validation one, and the errors indicate a drift in laboratory chemistry, instrument variations or wavelength instability [69].

Some of the symbols most widely used and applied in this dissertation are listed below.

Symbols used in formulae:

- ∑ Sum of all values in parenthesis
- N Total number of samples used
- ŷ Predicted values
- $\bar{Y}$ Mean value
- Y Laboratory values
- K Number of wavelengths used in an equation

1.3.4.1 *Coefficient of determination*

$$R^2 = \left(\frac{\sum_{i=1}^{N}(\hat{y}i - Y)^2}{\sum_{i=1}^{N}(Y - \bar{Y}i)^2}\right)$$

This statistic expresses the amount of variation in the data which is suitably modelled by the calibration equation as a total fraction of 1.0. If R^2 is equal to 1, the model explains 100% of the variation within the data. The lower the value obtained, the lower the amount of variance explained, and for that reason when performing a calibration model, we always aim to get an R^2 of 1.0.

1.3.4.2 Standard Error of Calibration (SEC)

$$SEC = \left(\frac{\sum_{i=1}^{N} (Yi - \hat{y})^2}{N - K - 1} \right)^{1/2}$$

This statistic is the Standard deviation for the residuals produced by the real or wet laboratory analytical values and the predicted NIR values for samples in the calibration set. The statistic SEC describes how samples fit into a calibration model.

1.3.4.3 Standard Error of Prediction (SEP)

$$SEP = \sqrt{\frac{\sum_{i=1}^{N} (Yi - \hat{y}i)^2}{N - 1}}$$

SEP shows the standard deviation for the residuals between real or wet analytical values and the predicted NIR values for samples outside the calibration set using a specific calibration equation. During the SEP calculation, the accuracy of the wet samples and the uniform distribution for the validation sample set are important.

1.3.4.4 Residual Predictive Deviation (RPD)

$$RPD = \frac{SD}{SEP}$$

RPD is the relationship between the Standard Deviation (SD) of the population reference values and the Standard Error of Prediction (SEP).

If RPD values are below 1.5, the model should be considered poor and not usable; when values of RPD lie between 1.5 and 2, the models can be considered fair, whereas when those values are between 2.5 and 3 or above 3, the predictions performed by the models can be considered good and excellent, respectively [75].

1.3.4.5 Ratio Error Range (RER)

$$RER = \frac{Range}{SD}$$

The values obtained with RER will normally be around four to five times greater, and more influenced by extreme results, than those with RPD. When the RER value is below 4, the calibration is acceptable for screening; between 10 and 15, the calibration is fairly good for quality control and above 15 is suitable for quantification.

Bibliography:

[1] J. F. Hancock. Plant evolution and the Origin of Crop Species. CABI publishing, Cambridge, MA, USA (2004).

[2] A. B. Smith. The Role of Food, Agriculture, Forestry and Fisheries in Human Nutrition. Vol. I- Animal Husbandry, Nomadic Breeding and Domestication of Animals. Edited by Victor R. Squires. EOLSS Publishers Co Ltd (2009).

[3] M. Zvelebil and M. Pluciennik. The Role of Food, Agriculture, Forestry and Fisheries in Human Nutrition. Vol. I- Historical Origins of Agriculure. Edited by Victor R. Squires. EOLSS Publishers Co Ltd (2009).

[4] P. J. Richerson, M. B. Mulder and B. Vila. Principles of Human Ecology. Chapter 3. Hunting and Gathering Societies. Pearson Custom Pub; Custom edition (1996).

[5] I. N. Edith and E. M. Ochubiojo. Scientific, Health and Social Aspects of the Food Industry. 21. Food Quality Control: History, Present and Future. Edited by Benjamin Valdez, InTech (2012).

[6] A. Miltra. Fundamentals of Quality Control and Improvement. John Wiley & Sons, (2012).

[7] R. Lásztity, M. Petró-Turza and T. Földesi. Food Quality and Standards. Vol I. History of the Food Quality Standards. EOLSS Publishers Co Ltd (2009).

[8] J. Slavin. Whole grains and human health. Nutrition Research Reviews (2004), 17, 99–110.

[9] L. Cordain. Cereal Grains: Humanity's Double-Edged Sword. World Rev Nutr Diet. Basel, Karger (1999), vol 84, pp 19–73.

[10] L. López Bellido. Cultivos Herbaceos Vol. I: Cereales. Ediciones Mundi-Prensa, (1990).

[11] R. Derpsch and T. Friedrich. Sustainable crop production intensification –The adoption of conservation agriculture worldwide-. 16th ISCO Congress, 8-12 Nov. (2010), Santiago, Chile.

[12] Food and Agriculture Organization (FAO), (2003). World agriculture: towards 2015/2030. (http://www.fao.org/index_en.htm).

[13] D. Southgate. Population Growth, Increases in Agricultural Production and Trends in Food Prices. *The Electronic Journal of Sustainable Development* (2009) 1(3).

[14] P. Kumar, R.K. Yadava, B. Gollen, S. Kumar, RK. Verma and S. Yadav. Nutritional Contents and Medicinal Properties of Wheat: A Review. Life Sciences and Medicine Research, Volume (2011): LSMR-22.

[15] Mateos-Aparicio I., Redondo Cuenca A., Villanueva-Suárez M. J. and Zapata-Revilla M. A. Soybean, a promising health source. Nutr Hosp. (2008); 23(4):305-312

[16] L. López Bellido. Cultivos industriales. Ediciones Mundi-Prensa, (2003).

[17] D. Zamora Zamora. Desarrollo de nuevas metodologías analíticas para el control de procesos y productos industriales: Aplicación de la Espectroscopía NIR y la Espectrometría de Movilidad Iónica (IMS). Doctoral thesis, University of Barcelona (2012).

[18] N. Sheppard. The Historical Development of Experimental Technique in Vibrational Spectroscopy. Handbook of Vibrational Spectroscopy, Vol. 5. Wiley (2001).

[19] W. Abeny and E. R. Festing. On the influence of the Atomic Grouping in the Molecules of Organic Bodies on Their Absorption in the Infra Red Region of the Spectrum. Philosophical Transactions of the Royal Society of London Vol. 172, (1881), pp. 887-918

[20] K. Wilbur. Near-Infrared Spectroscopy, A review. I. Spectral identification and analytical applications. Spectrcchimica Acts, 1954. Vol. 6. pp. 257 to 287.

[21] F.E. Barton II. Progress in near infrared spectroscopy the people, the instrumentation, the applications. In Near Infrared Spectroscopy: Proceedings of the 11th international near infrared spectroscopy conference, A.M.C. Davies and A. Garrido Varo (Eds.), NIR Publications, Chichester, West Sussex, UK, 2004, p. 13.

[22] P. H. Hindle. Handbook of Near-Infrared Analysis Third Edition. 1. Historical Development. Edited by Donald A. Burns and Emil W. Ciurczak. 2008 by Taylor & Francis Group, LLC.

[23] M.C. Sarraguca and J.A. Lopes. Quality control of pharmaceuticals with NIR: From lab to process line. Vibr. Spectrosc. 49 (2009) 204–210.

[24] M. Blanco and M. Alcala, Use of near-infrared spectroscopy for off-line measurements in the pharmaceutical industry. K.A. Bakeev (Ed.), Process Analytical Technology, Spectroscopic Tools and Implementation, Strategies for the Chemical and Pharmaceutical Industries. Blackwell Publishing, New York (2005) pp. 13–20, 362–376.

[25] J. Märk, M. Andre, M. Karner and C.W. Huck. Prospects for multivariate classification of a pharmaceutical intermediate with near-infrared spectroscopy as a process analytical technology (PAT) production control supplement. Eur. J. Pharm. Biopharm. 76 (2010) 320–327.

[26] M. Ito, T. Suzuki, S. Yada, H. Nakagami, H. Teramoto, E. Yonemochia and K. Terada. Development of a method for nondestructive NIR transmittance spectroscopic analysis of acetaminophen and caffeine anhydrate in intact bilayer tablets. J. Pharm. Biomed. Anal. 53 (2010) 396–402.

[27] S. Perrey. Non-invasive NIR spectroscopy of human brain function during exercise. Methods 45 (2008) 289–299.

[28] S. Kasemsumran, Y. P. Du, K. Murayama, M. Huehne and Y. Ozaki. Near-infrared spectroscopic determination of human serum albumin,γ-globulin, and glucose in a control serum solution with searching combination moving window partial least squares. Analytica Chimica Acta 512 (2004) 223–230.

[29] J. Wang, Y. J. Geng, B. Guo, T. Klima, B. N. Lal, J. T. Willerson and W. Casscells. Near-Infrared Spectroscopic Characterization of Human Advanced Atherosclerotic Plaques. Journal of the American College of Cardiology Vol. 39, No. 8 (2002).

[30] E. Cleve, E. Bach and E. Schollmeyer. Using chemometric methods and NIR spectrophotometry in the textile industry. Analytica Chimica Acta 420 (2000) 163–167.

[31] S. Macho and M. S. Larrechi. Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. trends in analytical chemistry, vol. 21, no. 12 (2002).

[32] A. M Bruno-Soares, I. Murray, R. M Paterson and J. M.F Abreu. Use of near infrared reflectance spectroscopy (NIRS) for the prediction of the chemical composition and nutritional attributes of green crop cereals. Animal Feed Science and Technology, Volume 75, Issue 1, 30 September (1998) Pages 15-25.

[33] O. K. Chung, J. B. Ohm, G. L. Lookhart and R. F. Bruns. Quality Characteristics of Hard Winter and Spring Wheats Grown under an Over-wintering Condition. Journal of Cereal Science, Volume 37, Issue 1, January (2003), Pages 91-99.

[34] B.R. Vazquez de Aldana, B. Garcia-Criado, A. Garcia-Ciudad and M. E. Perez-Corona. Non-destructive method for determining ash content in pasture samples: Application of near-infrared reflectance spectroscopy. *Commun.* Soil Sci. PlantAnal. (1996) 27(3-4):795-802.

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

[36] C. Petisco, B. García-Criado, B. R. Vázquez-de-Aldana, A. De Haro and A. García-Ciudad. Measurement of quality parameters in intact seeds of

Brassica species using visible and near-infrared spectroscopy. Industrial Crops and Products 32 (2010) 139–146.

[37] Z. Huang, S. Sha, Z. Rong, J. Chen, Q. He, D. M. Khan and S. Zhu. Feasibility study of near infrared spectroscopy with variable selection for non-destructive determination of quality parameters in shell-intact cottonseed. Industrial Crops and Products 43 (2013) 654–660.

[38] A. J. Gaitán-Jurado, M. García-Molina, F. Peña-Rodríguez and V. Ortiz-Somovilla. Near infrared applications in the quality control of seed cotton. *J. Near Infrared Spectrosc.* (2008) 16, 421–429.

[39] B. Kemps, L. Leon, S. Best, J. De Baerdemaeker and B. De Ketelaere. Assessment of the quality parameters in grapes using VIS/NIR spectroscopy. Biosystems engineering 105 (2010) 507–513.

[40] T. Nishizawa, Y. Mori, S. Fukushima, M. Natsuga and Y. Maruyama. Nondestructive analysis of soluble sugar components in strawberry fruits using near-infrared spectroscopy. Nippon Shokuhin Kagaku Kogaku Kaishi 56, 229–235. Journal of the Japanese Society for Food Science and Technology (2009) v. 56(4) p. 229-235.

[41] L. Yande, S. Xudong, Z. Jianmin, Z. Hailiang and Y. Chao. Linear and nonlinear multivariate regressions for determination sugar content of intact Gannan navel orange by Vis-NIR diffuse reflectance spectroscopy. Mathematical and Computer Modelling 51 (2010) 1438-1443.

[42] I. Kavdir, B. M. Burak, L. Renfu, K. Habib and S. Murat. Prediction of olive quality using FT-NIR spectroscopy in reflectance and transmittance modes. Biosystems Engineering 103 (2009) 304-312.

[43] C. Shiroma and L. Rodriguez-Saona. Application of NIR and MIR spectroscopy in quality control of potato chips. Journal of Food Composition and Analysis 22 (2009) 596–605.

[44] G. Macaloney, I. Draper, J. Preston, K. B. Anderson, M. J. Rollins, B. G. Thompson, J. W. Hall and B. McNeil. At-Line control and Fault Analysis in an Industrial High Cell Density *Escherichia Coli* Fermentation, Using NIR Spectroscopy. Food and Bioproducts Processing, Volume 74 (1996), Issue 4, 212-220

[45] D. Pérez-Marín, T. Fearn, J. Guerrero and A. Garrido-Varo. A methodology based on NIR-microscopy for the detection of animal protein by-products. Talanta 80 (2009) 48–53.

[46] S. F. Graham, S. A Haughey, R. M. Ervin, E. Cancouët, S. Bell and C. T. Elliott. The application of near-infrared (NIR) and Raman spectroscopy to detect adulteration of oil used in animal feed production. Food Chemistry 132 (2012) 1614–1619.

[47] J. Zhao, H. Lin, Q. Chen, X. Huang, Z. Sun and F. Zhou. Identification of egg's freshness using NIR and support vector data description. Journal of Food Engineering 98 (2010) 408–414.

[48] G. Tøgersen, T. Isaksson, B.N. Nilsen, E.A. Bakker and K.I. Hildruma. On-line NIR analysis of fat, water and protein in industrial scale ground meat batches. Meat Science 51 (1999) 97-102.

[49] J.A. Fernández Pierna, P. Vermeulen, O. Amand, A. Tossens, P. Dardenne and V. Baeten. NIR hyperspectral imaging spectroscopy and chemometrics for the detection of undesirable substances in food and feed. Chemometrics and Intelligent Laboratory Systems 117 (2012) 233–239.

[50] B. M. Nicolaï, E. Lötze, A. Peirs, N. Scheerlinck and K. I. Theron. Non-destructive measurement of bitter pit in apple fruit using NIR hyperspectral imaging. Postharvest Biology and Technology 40 (2006) 1–6.

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

[52] D. Pérez-Marín, T. Fearn, J. E. Guerrero, A. Garrido-Varo. A methodology based on NIR-microscopy for the detection of animal protein by-products. Talanta 80 (2009) 48–53.

[53] D. Pavino, S. Squadrone, M. Cocchi, G. Martra, D. Marchis and M.C. Abete. Towards a routine application of vibrational spectroscopy to the detection of bone fragments in feedingstuffs: Use and validation of a NIR scanning microscopy method. Food Chemistry 121 (2010) 826–831.

[54] F. J. Schmitt, H. Südmeyer, J. Börner, J. Löber, K. Olliges, K. Reineke, I. Kahlen, P. Hatti, H. J. Eichler and H. J. Cappius. Handheld device for fast and non-contact optical measurement of protein films on surfaces. Optics and Lasers in Engineering 49 (2011) 1294–1300.

[55] D. Pérez-Marín, P. Paz, J. E. Guerrero, A. Garrido-Varo and M. T. Sánchez. Miniature handheld NIR sensor for the on-site non-destructive assessment of post-harvest quality and refrigerated storage behavior in plums. Journal of Food Engineering 99 (2010) 294–302.

[56] S. Kawano. Progress in Application of NIR and FT-IR in Food Characterization. Characterization of Food: Emerging Methods. Edited by A. G. Gaonkar, (1995) Elsevier.

[57] G. Reich. Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. Advanced Drug Delivery Reviews 57 (2005) 1109– 1143.

[58] E.W. Ciurczak, Handbook of Near-Infrared Analysis, Burns D.A. and Ciurczak E.W. (eds.), Marcel Dekker, New York, Basel (2001) p. 7.

[59] D. J. Hayes. Analysis of Lignocellulosic Feedstocks for Biorefineries with a Focus on the Development of Near Infrared Spectroscopy as a Primary Analytical Tool. Doctoral thesis, University of Limerick (2011).

[60] J. S. Shenk, J. J. Workman and Westerhaus, M. O. Application of NIR spectroscopy to agricultural products, In: Handbook of Near-Infrared Analysis, D. A. Burns, E. W. Ciurczak (Eds.), (2008) pp. (347-386), Taylor & Francis Group.

[61] M. Bautista Mercader. Avances en la aplicación de la espectroscopia NIR en la industria farmacéutica. Introducción a PAT y técnicas de imagen. Doctoral thesis, University of Barcelona (2009).

[62] C. Burgess. Chapter 1: The Basics of Spectrophotometric Measurement. UV-Visible Spectrophotometry of Water and Wastewater. (2007) Elsevier.

[63] N. K. Afseth and A. Kholer. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. Chemometrics and Intelligent Laboratory Systems 117 (2012) 92–99.

[64] Y. Liu, W. Liu, X. Sun, R. Gao, Y. Pan and A. Ouyang. Potable NIR spectroscopy predicting soluble solids content of pears based on LEDs. Journal of Physics: Conference Series 277 (2011) 012026.

[65] B.G.M. Vandeginste and D.L. Massart. Handbook of chemometrics and qualimetrics. Elsevier Science, (1998).

[66] H.M. Heise and R. Winzen. Near-Infrared Spectroscopy. Principles, Instruments, Applications. Edited by H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise. Wiley-VCH, Weinheim, Germany, (2002).

[67] Y. Ozaki, S. Morita and Y. Du. Near-Infrared Spectroscopy in Food Science and Technology. Edited by Yukihiro Ozaki, W. fred McCLure and Alfred A. Christy. John Wiley & Sons Inc., Hoboken, New Jersey, USA, (2007).

[68] K. R. Beebe, R. J. Pell and M. B. Seasholtz. Chemometrics. A practical guide, John Wiley & sons, New York (1998).

[69] R. Riovanto. Near Infrared Spectroscopy in Food Analysis: Qualitative and Quantitative Approaches. Doctoral thesis (2011).

[70] V. Fernández Cabanas. Métodos de procesamiento de la señal espectroscópica NIRS: Aplicación al análisis Cuantitativo y Cualitativo de productos alimentarios. Tesis doctoral, Universidad de Córdoba (2003).

[71] H. Cen and Y. He. Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science & Technology 18 (2007) 72e83.

[72] D. Svozil, V. KvasniEka and J. Pospichal. Introduction to multi-layer feed-forward neural networks. Chemometrics and Intelligent Laboratory Systems 39 (1997) 43-62.

[73] H. Martens and T. Naes. *Multivariate calibration*, John Wiley & Sons, England (1989).

[74] M. Kenneth Boysworth and K. S. Booksh. Aspects of Multivariate Calibration Applied to Near-Infrared Spectroscopy. Handbook of Near-Infrared Analysis, third edition. Edited by Donald A. Burns and Emil S. Ciurczak. Taylor & Francis Group (2008).

[75] W. Saeys, A. M. Mouazen and H. Ramon. Potential for Onsite Analysis of Pig Manure using Visible and Near Infrared Reflectance Spectroscopy. Biosystems Engineering (2005) 91 (4), 393-402.

Chapter II

Development of robust soybean NIR calibration models with high variability and temperature compensation in the base data

2.1 Introduction

Soybean (*Glycine max*) is one of the most widely grown crops in United States and the world. Total production of 2011 was lead by USA (90,609,800 mt), Brazil (68,518,700 t) and Argentina (52677400 t) [1]. Therefore, the demand for soybean products has increased noticeably. Soybeans are main protein source for animal and human food as well as one of the main vegetable oils to produce biodiesel and other industrial uses as adhesives, coatings and printing inks, lubricants, plastics, etc. [2].

High-protein, high-oil or high-yield cultivars have increased soybean crop value. Over the years conventional or wet analytical methods have been used for the determination of the parameters desired. These analytical methods usually offer accurate primarily proximate results, but are time consuming and have the disadvantage of being impractical for repetitive measurements [3]. For that reason the food industry and nutritional sciences have need for rapid techniques that are economical, accurate, reproducible and non-destructive.

The acceptance of Near Infrared Reflectance Spectroscopy (NIRS) [4], [5] has arisen as an alternative for the determination of various constituents in different matrices [6]. This technology has already been accepted by the American Association for Clinical Chemistry (AACC), Association of Analytical Chemists (AOAC), International Association for Cereal Science and Technology (ICC), American Oil Chemists' Society (AOCS) and International Organization for Standardization (ISO) confirmed its applicability for routine use [7] (Table 1).

Table II-1. Official standard methods for quality measuring of parameters by NIR

Standardization organization	Analytical Method
AACC	*08-21.01* Ash Content in Wheat Flour *39-10.01* Protein Determination in Small Grains *39-20.01* Protein and Oil Determination in Soybean *39-70.02* Hardness Determination in Wheat
AOAC	*997.06* Protein (crude) in Wheat *2007.04* Moisture, Protein, and Fat in Meats
ICC	*159* Protein Determination in Ground Wheat and Flour
ISO	12099:2010 Determination of Moisture, Fat, Protein, Starch and Crude Fiber in Animal Feeding Stuffs, Cereals and Milled Cereal Products 21543:2006 Milk Products.
AOCS	AM-1.09a Guideline

NIRS provides high performance by increasing product quality determination and reducing analysis time and cost [8]. Multivariate statistics are used to develop quantitative models with the information from near infrared spectra. Instrument model-specific equations for physical, chemical and biological properties have been developed. The modelling process comprises the following steps: data exploration, data pre-processing, calibration design, modelling, validation and application.

Zeaiter [9] defined robustness as "the predictive capacity against perturbations centred on standard conditions". The lack of robustness in calibration models can be caused by variations in the experimental conditions. In looking for reliable prediction, variations which could cause loss of model robustness and prediction accuracy must be taken into account. Cozzolino [10], described sample temperature as the external factor most widely studied as affecting calibration robustness. Changes in sample temperature could trigger changes in the intensity of the molecular vibrations, resulting so, in changes within the spectra.

The basis of the quantitative analysis is establishing a relation between instrument outputs and the property desired**.** For calibration, it is necessary to have a previous knowledge of the variables that are going to be determined. Reference methods must provides the most precise and accurate values of the analyte.

2.2 Objectives

The general objective of this study was to ensure that the models developed were capable of passing the requirements of the National Type

Evaluation Program (NTEP). This program checks the effects of power supply fluctuations, storage temperature, levelling, warm-up time, humidity, instrument stability and instrument temperature sensitivity. Passage is necessary for use in trade in the United States. With the specific objectives of:

- Characterization and study of the diverse soybean spectral information coming from all over the world.
- Development of robust models that cover different instrumentation models within the manufacturers' line.
- Inclusion of spectra obtained at Cold, Room and Hot temperatures making the models more robust to temperature changes.

Including most of the variability possible in the calibration set, will make those capable of predicting with high accuracy constituents of future new samples.

2.3 Materials and Methods

2.3.1 Samples and spectra collection

In this study were used 4,179 protein soybeans samples and 4,183 oil ones from all world origins during 2001-2010 crop seasons. These samples were received, stored and scanned in the Grain Quality Laboratory, placed in food sciences building in Iowa State University. This laboratory provides

instrumental analyses of the chemical and physical properties of grain and other agricultural products.

Spectra were collected in transmittance mode (log 1/T) using four instruments: OmegAnalyzerG serials 106110, 106118 and 609448 and AgriCheck 31002 (Bruins Instrument, Puchheim, Germany). All are grating monochromator based units with a silicon detector (Figure II-1).

Figure II-1. Disposition of the units into the Grain Quality Laboratory

The spectrum range was from 850 to 1048 nm, every 2 nm interval, producing a total of 100 data points (Figure II-2).

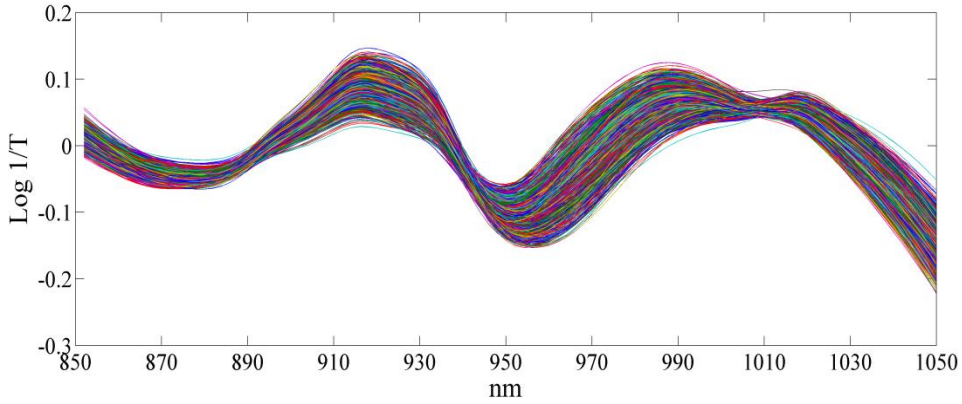


Figure II-2. Representation of raw spectra of soybean scanned by OmegAnalyzer G 106110

AgriCheck in approximately 40 seconds analysis time takes 10 subsamples, using a fixed path length cell (30 mm for soybean) while OmegAnalyzerG has 50 second analysis time for 16 subsamples, having a variable path length from 8 to 30 mm depending on the variety of grain and seed.

Sample scans were taken over all instruments at the same day, starting from OmegAnlyzer G 609448 around the laboratory room.

2.3.2 Reference values

Protein was determined by combustion following the standard of AOAC (990.03). Nitrogen freed by combustion at high temperature in pure oxygen is measured by thermal conductivity detection, and converted to equivalent protein percentage, by appropriate numerical factor.

Oil was determined by ether extract (AOCS Ac 3-44). This method determines the substances extracted from ground soybean seeds by petroleum

ether under the conditions of the test. Approximately 0.6-0.7% of the soybean oil extracted is phospholipids.

Over the 10 years of the spectra collection, all the reference data came from the same laboratory: Eurofins Scientific, Des Moines, Iowa (http://www.eurofinsus.com/).

2.3.3 Calibration and validation set

Two set of samples were required, the first one for construction the library and the second as an independent group for validating the model [11]. Two major conditions have to be achieved: 1) representativeness and diversity of both calibration and validation sets and 2) statistical independence of the validation set [12].

The starting calibration group (4,179 items for protein and 4,183 for oil) contained the spectra of samples coming from nine crop seasons (2001-2009) scanned by the four NIR spectrometers.

As external validation an independent group formed with the spectra of soybean samples from the 2010 crop season was used. In this case, only three instruments (106110, 106118 and 31002) were used, for technical problems. This validation set configuration created environmental true validation because the validation samples were not statistically or related to the calibration set.

2.3.4 Data analysis

Two software were used: 1) the Unscrambler 9.8 (Camo a/s Oslo, Norway) for calculations and spectral management and 2) for data selection, Matlab version 7.0 R-14 (The MathWorks, Natick, MA, USA).

2.3.4.1 Data selection of the calibration collective

The procedure of this work was selecting an appropriate calibration set. The selection was based on splitting the matrices into blocks of related variables, creating, in this way, a hierarchical model [13]. The blocks, in this case, were defined on instrument per year; randomly it was selected one year coming from each instrument, ensuring the total variability included in the data set: years and instrument (Figure II-3). Blocking the information generally simplifies the interpretation. Therefore, the initial sets were reduced to 1,095 samples for protein and 1,072 for oil.

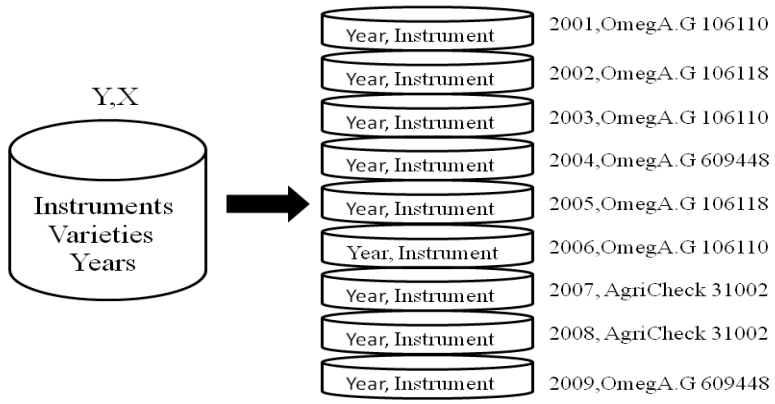


Figure II-3. Representation of the hierarchical model based on years and instrumentation.

2.3.4.2 Sample Temperature spectra acquisition

Additional samples scanned in OmegAnalyzerG 106110 and 106118 at different temperatures were expected to make the model insensitive to temperature variations [14] and [15].

The procedure to follow was established by the Grain Quality Laboratory (Iowa State University, Iowa, USA): 1) samples were run in the temperature order: cold (5ºC), room (22ºC), warm (45ºC); 2) samples were always run in ascending numeric order, following a pre-printed list of sample ids; 3) sample quantity was large enough to have enough for one complete fill of each instrument, avoiding refills and recycling (not allowed); 4) samples was sealed at all times except during use; 5) all samples were run at the beginning, at room temperature, in a capacitance moisture to monitor for moisture changes.

The process started analyzing cold samples which have been previously maintained in the cooler during 24 hours. After running cold samples they were placed into lab conditions another 24 hours, to temper the sample to RT before scanning them again.

Once the samples were scanned at CT and RT they passed to be warmed. Using a Gamet rotating divider (seedburo, Ins., Chicago, IL), about 500 grams of each sample were placed into an oven for approximately 2 hours to reach the temperature of 45ºC and they were scanned immediately (Figure II-4).

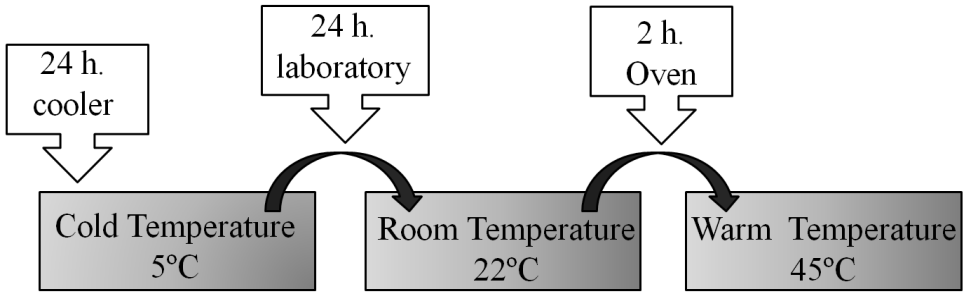


Figure II-4. Flow chart of temperature compensation spectra collection.

A total of 104 samples were run at cold temperature (CT), room temperature (RT) and warm temperature (WT). Following the procedure explained before.

2.3.4.3 Principal component analysis

PCA (described in section 1.3.3.1.1) was carried out as an explorative analysis of the data structure for outlier identification. The score algorithm ranks spectra according to Mahalanobis distance (MD), depending on the spectral and chemical variability of samples in the population.

The application of this algorithm in order to reduce the number of variables in the correlated dataset [16] was used on the calibration set, leaving out the validation group so as to simulate a real situation. The spectral pre-treatment for PCA were the same as the ones in the regression analysis. The criterion to selected sample as anomalous was marked as outlier in the 70% of the analysis (5 of 7 analyses).

2.3.4.4 Regression models

Partial Least Squares plus cross validation (describe in section 1.3.3.2.1) was used for model development.

A total of 9 combinations as a result of different pre-treatment and derivative were used. So that, the first combination used was raw spectra, raw + first derivative and raw + second derivative. For the spectral pre-treatment Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC) (previously described in section 1.3.2.4) the same combination was used.

The statistic parameters used to evaluate the models were: The coefficient of determination (R^2), the Standard Error of Prediction (SEP), the Ratio of Standard Error of Prediction to Standard Deviation (RPD), the Ratio Error Range (RER). These statistical analyses were explained in the introduction of the thesis in the section 1.3.4.

2.4 Results and Discussion

2.4.1. Reference analyses

Before building the model, a study of the reference data was done. This stage is important in order to know how the sample set is distributed. This allowed to select unusual samples and included them into the calibration model so as to increase the collective range.

Statistical analyses of the mean, standard deviation, maximum, minimum and range over the different years was done before starting with the multivariate analyses. Table 2 and 3 show the reference values and their statistical analyses.

Table II-2. Results of the reference values and statistical analysis of protein

	Years									*Ext.Val.*
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Mean	36.02	37.32	36.98	36.32	36.00	36.44	37.39	35.43	35.49	36.01
SD	2.54	3.41	2.49	2.97	3.89	4.34	3.20	4.08	4.29	4.22
Max	44.18	46.50	44.00	45.07	46.04	46.31	45.32	44.41	46.89	47.17
Min	31.28	30.11	30.63	27.48	29.74	28.80	30.95	24.72	25.26	28.91
Range	12.90	16.39	13.37	17.59	16.30	17.51	14.37	19.69	21.63	18.26

Table II-3. Results of the reference values and statistical analysis of oil

	Years									*Ext.Val.*
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Mean	18.11	18.57	17.76	18.15	18.85	18.23	18.10	18.43	18.50	18.25
SD	1.34	1.70	1.56	1.68	2.10	2.21	1.91	1.70	2.79	1.74
Max	21.58	21.85	21.02	21.94	22.52	23.57	22.13	22.36	24.64	21.88
Min	14.35	13.35	12.48	12.74	13.51	12.56	13.70	12.49	11.85	13.87
Range	7.23	8.50	8.54	9.20	9.02	11.00	8.43	8.87	12.79	8.01

As Earl [17] points out, oil content in soybeans tends to be negatively correlated with protein content. Tables 1 and 2 showed years such as 2002, 2003 and 2007 present higher values in protein with lower percentages in oil. On the contrary, years as 2005, 2008 and 2009 present the lesser content in protein but the maximum in oil.

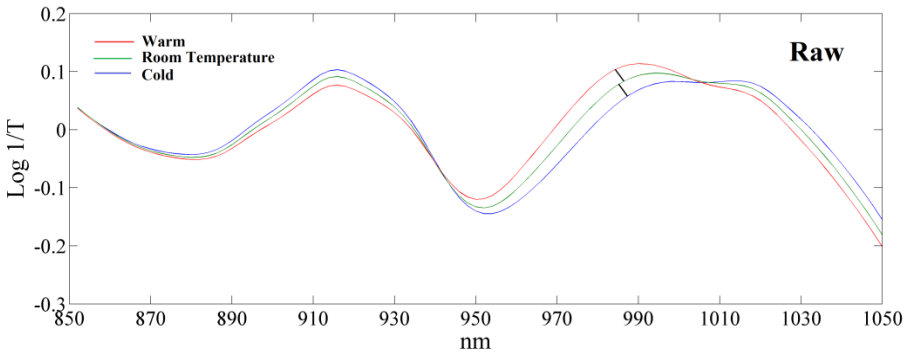
For model building, some extreme samples were removed taking into consideration that typical composition of soybeans seeds is 40.69±0.51 for protein content and 21.38±0.64 for oil content [17] but maintaining the

widest possible range for both constituents in order to cover future variations on the composition.

2.4.2 Soybean seed spectra with temperature compensation

Since absorption bands provide information about the individual bonds and the interaction among molecules, usually some interactions (such as hydrogen bonding) are very weak to be broken with the increase of the temperature, causing shifts on the spectral profiles.

Important variation was observed in the NIR spectra of soybean seeds analyzed at different temperatures. Figure II-5 displays the raw and second derivative spectra of the mean spectra corresponding to cold, room and warm temperatures.



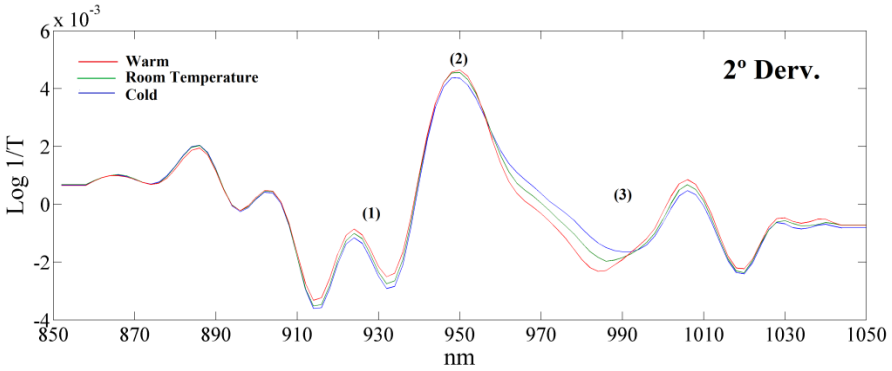


Figure II-5. Near infrared mean spectrum of temperature samples. On Raw spectra and after Second derivative.

Visually, as it was expected, the three spectra followed a similar pattern on the absorption bands situated in the same wavelength. However, in certain wavelengths, there were differences in shifts in the absorbance levels. When second derivative was applied, the intensity dissimilarities of the absorption peaks were remarkable.

Absorption bands around 950 nm and 990 nm are most influenced by water (O-H second overtone). It is visible peak at 950 nm which corresponds to oil absorption [18]. Changes related to the temperature modification around the O-H bonds in NIR region are clearly observed. As Wülfert [19] pointed out, the influence of the sample temperature produces deviation from linearity and additive effects, perfectly visible at 930 (1), 950 (2) and 990 (3) nm peaks.

2.4.3 Principal Component Analysis

2.4.3.1 Sample collective

The principal component analysis was performed first with raw spectra and then with pre-treatment, in order to visualize the tendency of the calibration set.

Figure II-6 displays: A) the score plot of the three principal components coming from the analyses of the 9 crop season on raw spectra and B) after Standard Normal Variate (SNV) + 2º derivative.

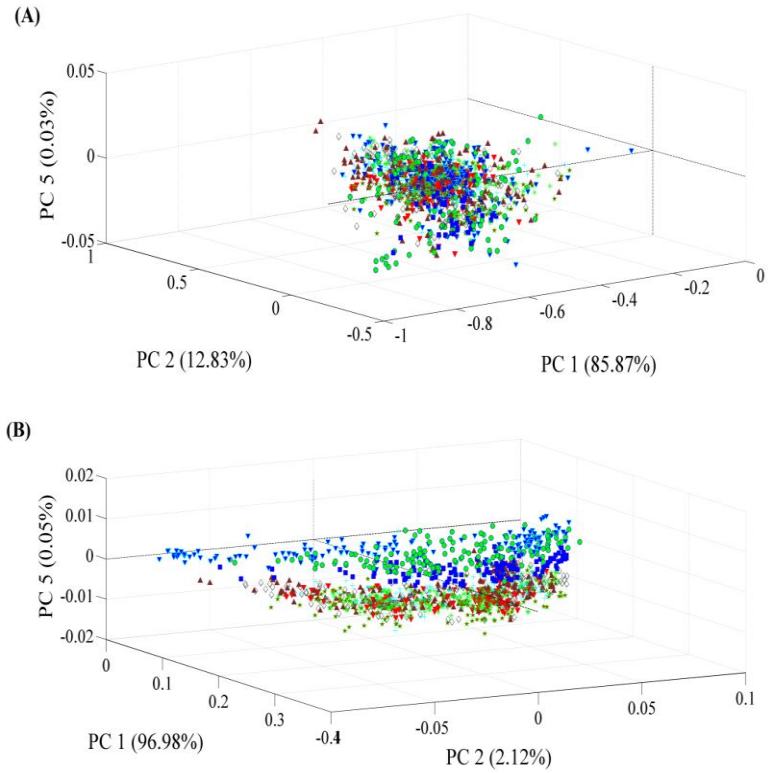


Figure II-6. PC1, PC2 and PC5 scores representation of the distribution during the years 2001-2009. (A) raw spectra and (B) SNV+ 2º derivative.

With raw spectra (Figure 3.A) it cannot be distinguished seasonal differences. Scores representation displays a compact clustering of total amount of samples. Within the first PCs most of the information is explained, only the first PC accounted for 85.88% of the total variance, with second one an additional 12.82%. With these two components was explained more than three quarters of the total variance, and even the fifth component was still making a modest contribution (0.03%).

On the other hand, when spectral pre-treatment was applied (Figure 3.B) the contribution of the first PC increased 85.87% to 96.98%. This reflects that the signal of derivative spectra enhance and improve the relation signal to noise ratio, once the noise and the redundant information was extracted [20]. Observing the PCA representation of PC1, PC2 and PC5 it can be distinguished a cluster formed with three of the nine years, which corresponds to 2003, 2008 and 2009. 2008 and 2009 form two of the seasons which have less quantity in protein. On the contrary, 2003 represent the year with lower mean value of oil. The 2009 cluster was most likely caused by rainfall during the season producing soybean seeds with high moisture.

After outlier detection, the calibration set was reduced into 1,059 spectra for protein and 963 for oil.

2.3.4.2 Sample temperature

As it was said in the section 2.3.2, depending on the temperature the spectra presented differences in the absorption peaks. Figure II-7 shows the scores plot corresponding to PC2, PC3 and PC4 axis. There is an obvious

clustering of three well defined groups (corresponding to cold, warm and room temperature samples).

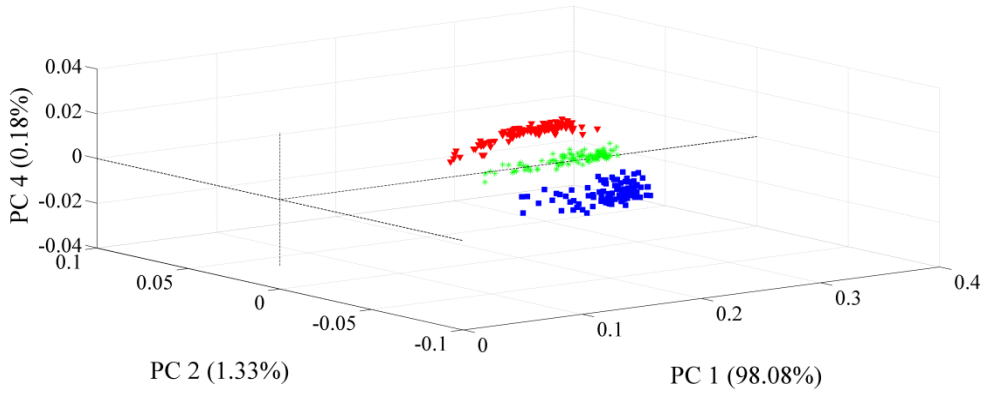


Figure II-7. PC1, PC2 and PC4 representation of the temperature samples (Blue: cold, Green: room temperature and Red: warm).

Even though 99.39% of the variance is explained with the first two PCs, the information provided by PC4 was highly important since it helps in distinguishing the groups based on different sample temperature spectra.

2.4.4 Partial Least Square modelling

The final models including temperature had 1,359 (after taking out two outliers coming from temperature compensation samples) and 1,275 samples for protein and oil respectively. As it was told in the section 2.2.4.4 the development of the models was done with 7 pre-treatment.

Tables II-3 and II-4 show the comparative results of the calibration with no temperature samples included (A) and with them into the calibration data set (B).

Table II-4. A) Results of protein calibration without temperature compensation

Calibration		Pretreatment						
		Raw	*Raw + 1st derv*	*Raw + 2nd derv*	*SNV + 1st derv*	*SNV + 2nd derv*	*MSC + 1st derv*	*MSC + 2nd derv*
	samples	1059	1059	1059	1059	1059	1059	1059
	R^2	0.96	0.96	0.96	0.97	0.97	0.98	0.97
	SEC	0.66	0.66	0.65	0.58	0.57	0.57	0.56
External validation								
OmegAnalyzer G 106110	*Samples*	138	138	138	138	138	138	138
	R^2	0.93	0.94	0.94	0.96	0.96	0.95	0.92
	SEP	0.91	0.90	0.88	0.63	0.61	0.61	0.65
OmegAnalyzer G 106118	*Samples*	138	138	138	138	138	138	138
	R^2	0.95	0.96	0.96	0.98	0.98	0.97	0.96
	SEP	0.87	0.86	0.83	0.59	0.58	0.56	0.66
AgriCheck 31002	*Samples*	134	134	134	134	134	134	134
	R^2	0.93	0.93	0.94	0.96	0.96	0.94	0.87
	SEP	0.89	0.90	0.89	0.64	0.69	0.60	0.64

Table II-4. B) Results of protein calibration including temperature compensation

Temperature Compensation		Pretreatment						
Calibration		*Raw*	*Raw + 1 st derv*	*Raw + 2 nd derv*	*SNV + 1 st derv*	*SNV + 2 nd derv*	*MSC + 1 st derv*	*MSC + 2 nd derv*
	samples	1369	1369	1369	1369	1369	1369	1369
	R^2	0.96	0.96	0.96	0.97	0.97	0.97	0.97
	SEC	0.62	0.61	0.60	0.53	0.52	0.54	0.52
External Validation								
OmegAnalyzer G 106110	*Samples*	138	138	138	138	138	138	138
	R^2	0.94	0.94	0.94	0.96	0.96	0.91	0.94
	SEP	0.90	0.89	0.88	0.62	0.62	0.68	0.65
OmegAnalyzer G 106118	*Samples*	138	138	138	138	138	138	138
	R^2	0.95	0.96	0.96	0.98	0.98	0.98	0.97
	SEP	0.86	0.85	0.84	0.57	0.58	0.58	0.69
AgriCheck 31002	*Samples*	134	134	134	134	134	134	134
	R^2	0.94	0.93	0.94	0.96	0.96	0.93	0.91
	SEP	0.90	0.89	0.89	0.64	0.66	0.56	0.63

Table II-5. A) Results of oil calibration without temperature compensation

Calibration		Pretreatment						
		Raw	*Raw + 1st derv*	*Raw + 2nd derv*	*SNV + 1st derv*	*SNV + 2nd derv*	*MSC + 1st derv*	*MSC + 2nd derv*
	samples	963	963	963	963	963	963	963
	R^2	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	SEC	0.39	0.4	0.39	0.4	0.38	0.39	0.38
External validation								
OmegAnalyzer G 106110	*Samples*	136	136	136	136	136	136	136
	R^2	0.93	0.93	0.93	0.94	0.94	0.91	0.94
	SEP	0.52	0.53	0.52	0.48	0.48	0.54	0.5
OmegAnalyzer G 106118	*Samples*	136	136	136	136	136	136	136
	R^2	0.89	0.88	0.76	0.91	0.81	0.88	0.78
	SEP	0.52	0.57	0.56	0.54	0.52	0.6	0.54
AgriCheck 31002	*Samples*	133	133	133	133	133	133	133
	R^2	0.92	0.93	0.91	0.94	0.92	0.91	0.92
	SEP	0.51	0.51	0.55	0.49	0.53	0.53	0.54

Table II-5. B) Results of oil calibration including temperature compensation

Temperature Compensation		Pretreatment						
Calibration		*Raw*	*Raw + 1st derv*	*Raw + 2nd derv*	*SNV + 1st derv*	*SNV + 2nd derv*	*MSC + 1st derv*	*MSC + 2nd derv*
	Sample	1275	1275	1275	1275	1275	1275	1275
	R^2	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	SEC	0.37	0.37	0.37	0.36	0.35	0.36	0.35
External Validation								
OmegAnalyzer G 106110	R^2	0.93	0.93	0.93	0.94	0.94	0.92	0.94
	RMSEP	0.53	0.52	0.52	0.48	0.48	0.55	0.49
	SEP	0.52	0.52	0.52	0.48	0.48	0.53	0.49
OmegAnalyzer G 106118	R^2	0.88	0.85	0.78	0.86	0.82	0.91	0.91
	RMSEP	0.69	0.77	0.92	0.74	0.84	0.56	0.57
	SEP	0.55	0.55	0.54	0.52	0.52	0.58	0.57
AgriCheck 31002	R^2	0.92	0.92	0.91	0.93	0.92	0.91	0.91
	RMSEP	0.52	0.54	0.56	0.49	0.53	0.56	0.57
	SEP	0.52	0.53	0.56	0.49	0.53	0.51	0.56

Regression technique based on NIR display the stronger ability to predict protein and oil parameters based on wet chemistry data.

With raw spectra, even applying derivative pre-treatment in protein the SEP values are high, certainly caused by the difficulty in spectral band identification because they are less intensive and noisier [21]. On the contrary, oil differences when applying pre-treatment and raw spectra are not very significant.

Best results obtained with the lowest SEP, in both case protein and oil, were those obtained with SNV + 2^{nd} derivative. When including temperature compensation samples, there were no significant changes.

Williams, 2001[22], expressed the guidelines for the interpretation of the RPD showed in Table II-5.

Table II-6. Guidelines of the RPD values

RPD	Classification	Application
0.0-2.3	Very poor	Not recommended
2.4-3.0	Poor	Rough screening
3.1-4.9	Fair	Screening
5.0-6.4	Good	Quality Control
6.5-8.0	Very good	Process control
8.1+	Excellent	Any application

Daniel [23] advised the limits of the values for RER in lignocellulosic feedstock in model performance: When RER is lower than 4, the calibration

could be acceptable for screening; between 10 and 15, the calibration should be fair for quality control and higher than 15 suitable for quantification.

Armstrong 2006 [24], reported RPD of 4.9 for protein content in a single soybean kernel; Wang 2012 [25] determined a RPD of 6.53 for peanut. Esteve 2012 [26], who compared instrumental response, reported RPD values of whole soybean seed between 1.46 and 4.46. According to Williams 2001 [22] prediction models for oil are proved to be acceptable for use as screening method. RPD values between 4.18 (OmegA.G 106110) and 3.86 (AgriCheck 31002) are obtained with SNV+1st derivative.

RPD and RER values obtained in both models are expressed in table II-6 and II-7 for protein and oil, respectively.

Table II-7. RPD and RER values of protein

		Raw	*Raw + 1 st derv*	*Raw + 2 nd derv*	*SNV + 1 st derv*	*SNV + 2 nd derv*	*MSC + 1 st derv*	*MSC + 2 nd derv*
NO TC	RPD_{106110}	4.53	4.58	4.69	6.55	6.76	6.76	6.34
	RER_{106110}	16.01	16.19	16.56	23.13	23.89	23.89	22.42
	RPD_{106118}	4.74	4.80	4.97	6.99	7.11	7.36	6.25
	$RER_{1061118}$	16.75	16.94	17.55	24.69	25.12	26.02	22.08
	RPD_{31002}	4.63	4.58	4.63	6.44	5.98	6.87	6.44
	RER_{31002}	16.37	16.19	16.37	22.77	21.12	24.28	22.77
TC	RPD_{106110}	4.58	4.63	4.69	6.65	6.65	6.06	6.34
	RER_{106110}	16.19	16.37	16.56	23.50	23.50	21.43	22.42
	RPD_{106118}	4.80	4.85	4.91	7.24	7.11	7.11	5.98
	RER_{106118}	16.94	17.14	17.35	25.56	25.12	25.12	21.12
	RPD_{31002}	4.58	4.63	4.63	6.44	6.25	7.36	6.55
	RER_{31002}	16.19	16.37	16.37	22.77	22.08	26.02	23.13

Table II-8. RPD and RER values of oil

		Raw	*Raw + 1st derv*	*Raw + 2nd derv*	*SNV + 1st derv*	*SNV + 2nd derv*	*MSC + 1st derv*	*MSC + 2nd derv*
NO TC	RPD_{106110}	3.86	3.78	3.86	4.18	4.18	3.71	4.01
	RER_{106110}	17.56	17.23	17.56	19.02	19.02	16.91	18.26
	RPD_{106118}	3.86	3.52	3.58	3.71	3.86	3.34	3.71
	$RER_{1061118}$	17.56	16.02	16.31	16.91	17.56	15.22	16.91
	RPD_{31002}	3.93	3.93	3.64	4.09	3.78	3.78	3.71
	RER_{31002}	17.90	17.90	16.60	18.64	17.23	17.23	16.91
TC	RPD_{106110}	3.86	3.86	3.86	4.18	4.18	3.78	4.09
	RER_{106110}	17.56	17.56	17.56	19.02	19.02	17.23	18.64
	RPD_{106118}	3.64	3.64	3.71	3.86	3.86	3.46	3.52
	RER_{106118}	16.60	16.60	16.91	17.56	17.56	15.74	16.02
	RPD_{31002}	3.86	3.78	3.58	4.09	3.78	3.93	3.58
	RER_{31002}	17.56	17.23	16.31	18.64	17.23	17.90	16.31

Regarding the results of RPD and RER, show that NIR models developed are suitable for prediction of protein and oil in whole soybean. Protein best values were obtained with models including temperature compensation into calibration set. The very high RPD of 7.11 and 7.36 and RER of 25.12 and 26.02 obtained with MSC+1derivative in OmegA.G 106118 and AgriCheck 31002 predictions, respectively; prove the suitability of the model for process control. However, OmegA.G 106110 best results obtained were when SNV spectral pre-treatment was used.

Oil displayed RPD and RER values lower than protein caused by the standard deviation of the sample set. It can be observed slightly differences between calibrations with temperature compensation and without it. The spectral pre-treatment together with 1 derivative showed the best results of the predictions in the three instruments. RPD values of 4.18 and 3.86 establish these models suitable for screening.

Bibliography

[1] http://www.fao.org/corp/statistics/es/

[2] http://www.soynewuses.org/

[3] C. Myoung-Gun. Determination of Sucrose Content in Soybean Using Near-Infrared Reflectance Spectroscopy. J. Korean Soc. Appl. Biol. Chem. 53(4), 478-484 (2010).

[4] D. L. B. Wetzel. Analytical near infrared spectroscopy. Instrumental Methods in Food and Beverage Analysis. (1998) Elsevier Science B.V.

[5] C. Haiyan and H. Yong. Theory and application of near infrared reflectance spectroscopy in determination of food quality. Trends in Food Science & Technology 18 (2007) 72-83.

[6] C. R. Bull. A review of sensing techniques which could be used to generate images of agricultural and food materials. Computers and Electronics in Agriculture,(1993) 8 1-29.

[7] M. Pojić, J. Mastilović and N. Majcen. Robustness of the near infrared spectroscopy method determined using univariate and multivariate approach. Food Chemistry 134 (2012) 1699-1705.

[8] F. Gogé, R. Joffre, C. Jolivet, I. Ross and L. Ranjanrd. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometrics and Intelligent Laboratory Systems 110 (2012) 168-176.

[9] M. Zeaiter, J.-M. Roger, V. Bellon-Maurel and D.N. Rutledge. Robustness of models developed by multivariate calibration. Part I: The assessment of robustness. Trends in Analytica Chemistry, (2004) Vol. 23, Nº 2.

[10] D. Cozzolino, L. Liu, W.U. Cynkar, R.G. Dambergs, L. Janik, C.B. Colby and M. Gishen. Effect of temperature variations on the visible and near infrared spectra of wine and the consequences on the partial least square calibrations developed to measure chemical composition. Analytical Chimica Acta 588(2007) 224-230.

[11] N. Broad, P. Graham, P. Hailey, A. Hardy, S. Holland, S. Hughes, D. Lee, K. Prebble and N. Salton and Paul Warren. Guidelines for the Development and Validation of Near-infrared Spectroscopic Methods in the Pharmaceutical Industry. Willey and sons (2002).

[12] F. Gogé, R. Joffre, C. Jolivet, I. Ross and L. Ranjanrd. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometrics and Intelligent Laboratory Systems 110 (2012) 168-176

[13] S. Wold, J. Trygg, A. Berglund and H. Antti. Some recent developments in PLS modeling. Chemometrics and Intelligent Laboratory Systems 58 (2001) 131-150.

[14] F. Wülfert, W. Th. Kok, O. E. de Noord and A. K. Smilde. Linear techniques to correct for temperature-induced spectral variation in multivariate calibration. Chemometrics and Intelligent Laboratory System 51(2000) 189-200.

[15] H. Swierenga, F. Wülfert, O.E. de Noord, A.P. De Weijer, A.K. Smilde and L.M.C. Buydens. Development of robust calibration models in near infra-red spectrometric applications. Analytica Chimica Acta 411 (2000) 121–135.

[16] W. G. Glen, W. J. Dunn III, and D. R. Scott. Principal Components Analysis and Partial Least Squares Regression. Tetrahedron Computer Methodology, Vol. 2 (1989) No. 6. Pp.349 to 376.

[17] E. G. Hammond, L. A. Johnson, C. Su, T. Wang and P. J. White. Bailey's Industrial Oil and Fat Products, Sixth Edition, Six Volumen Set. (2005) Jonh Wiley & Sons, Inc.

[18] A. G. Patil, M. D. Oak, S. P. Taware, S. A. Tamhankar and V. S. Rao Nondestructive estimation of fatty acid composition in soybean [*Glycine max* (L.) Merrill] seeds using Near-Infrared Transmittance Spectroscopy. Food Chemistry Volume 120,(2010), Pages 1210–1217.

[19] F. Wülfert, W. Th. Kok, and A. K. Smilde. Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models. Analytical Chemistry (1998) 70 (9), pp 1761-1767.

[20] A. D. Richardson, J. B. Reeves III and T. G. Gregoire. Multivariate analyses of visible/near infrared (VIS/NIR) absorbance spectra reveal underlying spectral differences among dried, ground conifer needle samples from different growth environments. New Phytologist (2003) 161: 291–301.

[21] T. Azzouz, A. Puigdoménech, M. Aragay and R. Tauler. Comparison between different data pre-treatment methods in the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-

squares multivariate calibration method. Analytica Chimica Acta 484 (2003) 121–134.

[22] P.C. Williams. Implementation of near-infrared technology. In: *Near*-Infrared Technology in the Agricultural and Food Industries (edited by P.C. Williams & K.H. Norris). (2001) Pp. 145-169. St. Paul, Minnesota, USA: St. Paul, Minnesota, USA.

[23] D. J. Hayes. Analysis of Lignocellulosic Feedstocks for Biorefineries with a Focus on The Development of Near Infrared Spectroscopy as a Primary Analytical Tool. Doctoral thesis (2011).

[24] P. R. Armstrong. Rapid Single-Kernel NIR measurement of grain and oilseed attributes. Applied Engineering in Agriculture Vol. 22(5): (2006) 767-772.

[25] L. Wang, Q. Wang, H. Liu, L. Liu and Y. Du. Determining the contents of protein and amino acids in peanuts using near-infrared reflectance spectroscopy. Journal of the Science of Food and Agriculture Volume 93 (2013), 118–124, 15.

[26] L. Esteve Agelet, P.R. Armstrong, C. I. Romagosa, and C. Hurburgh. Measurement of Single Soybean Seed Attributes by Near-Infrared Technologies. A Comparative Study. J. Agric. Food Chem. (2012) 60, 8314−8322.

Chapter IV

Application of Near Infrared Spectroscopy technology and hyperspectral NIR imaging for the detection of fungicide treatment on durum wheat samples

4.1 Introduction

Durum wheat is an important crop in the Mediterranean area. The main uses are in human food products, like bread, pasta and couscous [1]. The common Agricultural Policy Andalusia is the leading region producing durum wheat in Spain. It contributes more than 74% of the total national production [2].

Wheat genotypes, agronomics conditions and fertility inputs are the foremost factors determining durum wheat yield and quality characteristics [3]. Nevertheless, an important bounding aspect of durum wheat is the damage caused by diseases. Two of the most important are, above all, leaf rust and septoria leaf spot (incited by *Puccinia triticina* and *Septoria tritici*, respectively) (Figure IV-1).

Plant diseases are greatly influenced by environmental factors, including known stresses as deficiencies of essential nutrients and/or toxicities of other mineral elements [4]. Modifications in cultural practices, such as direct sowing, use of Nitrogen fertilizers and irrigation, may contribute to an increase on the disease severity [5]. *Puccinia triticina* is the most common rust of wheat. It has affected wheat for thousands of years. Yield losses in wheat from *Puccinia triticina* infections are usually the result of decreased numbers of kernels per head and lower kernel weights [6].

Figure IV-1. A) Puccinia Triticina, Ida Paul, Small Grain Institute, Bugwood.org. B) Septoria Tritici leaf disease of wheat. Clemson University - USDA Cooperative Extension Slide Series, Bugwood.org.

Methods used to fight fungal diseases and the development of new fungicides in cereals, are based on etiological and epidemiological knowledge. The presence of a particular fungal disease is related to the degree of susceptibility of the variety, presence of inoculum, plant phenological status and climatological factors, especially those associated with humidity [7].

When infective fungus is accumulated on the grain surface, enzymes destroy proteins, starch granules and grain cell walls [8], important for getting the standards. It reduces the yield and quality of grain and forage.

Seed coming from crops damaged by this disease have low vigour and thus poor emergence after germination. This can cause 100% yield losses if infection occurs very early and the disease keeps developing during the growing season [9].

Consumers are conscious of eating high quality products free of toxic agents. Increased food scrutiny requires the development of improved and more readily available analytical methods for food products authentication and detection of contaminant [10-13]. They want more information on dough and flour properties and are concerned about the variability in quality, both within and among lots Near Infrared Spectroscopy (NIRS) has been used for the determination and quantification of proximate quality parameters on food (protein, fat, sugar) and for the recognition of transgenic foods [14]. The basic assumption behind the application of spectroscopy to authentification lies with the generation of a "fingerprint" of foods. Given that each ingredient has a characteristic chemical composition it will also have a distinctive spectrum [15].

In the past decade, significant progress has been made in applying hyperspectral imaging technology in such applications as astronomy, remote sensing, and medical science [16]. The need for fast and reliable methods of authenticity and object identification has increased the interest in the application of hyperspectral near infrared imaging for quality control in the agricultural, pharmaceutical and food industries [17]. Early detection of apple bruises [18]; quality attributes in strawberry [19]; mushrooms [20]; [21];

bananas [22] and compound feed [23], among others has made that Hyperspectral imaging has emerged as a powerful technology.

Hyperspectral imaging technology integrates both imaging and spectroscopy into unique imaging sensors [24]. This technique provides the spectral information of a scanned sample, resulting in a spectrum per pixel.

The hyperspectral images can be described by a three dimensional array of size m _ n _ l, where m and n are the spatial dimensions (detector size or pixels) in the x and y directions and l is the wavelength or third dimension [25]. For this reason, Hyperspectral imaging presents a wide collection of data stored in pixels, which with the aid of multivariate analysis techniques, is capable of extracting the relevant information.

4.2 Objetives

The goal of this work was to evaluate the capability of NIR technology and hyperspectral NIR imaging to detect differences between durum wheat seed samples coming from plants which have been treated with fungicide and those coming from non treated plants, using discrimination models.

4.3 Materials and Methods

4.3.1 Experimental design

All the durum wheat samples used in this study came from trials carried out on randomised complete block designs with four replications

(Figure IV -2). Block design implies that there is only one observation for each treatment. It is the most common design used in field trials. Blocking removes as much variability as possible from the random error so that the differences among the groups are more evident. Crop management on trials was the standard used by farmers on the area. Experimental plots were $12m^2$ (10m x 1.20m).

Figure IV-2. Distribution of the blocks design in a field trial.

Although blocks are used in a randomized block design, the focus of the analysis is on the differences among the different groups.

4.3.2 Wheat samples

For the study were used a total of 213 wheat samples from 27 durum wheat varieties provided by the Andalusia Network of Agrarian Experimentation (RAEA), managed by the Instituto de Investigación y Formación Agraria y Pesquera (IFAPA). Samples originated from four different trial sites located in Jerez (Cadiz), Camino de Purchil (Granada),

Tomejil (Seville) and Santaella (Cordoba), each having different agroclimatic conditions. The 2009-2010 crop seasons were used.

For NIR analysis, IFAPA Center Alameda del Obispo (Córdoba), received the durum wheat seeds in paper bags containing approximately 500g, and then they were kept on small 250g plastic containers. Two samples of every variety were received, one, coming from fungicide treated plants (T) and the other from plants free of it (O), having a final set of 105 and 108 samples of each respectively.

For Hyperspectral imaging, the sample set was sent in plastic bottles of approximately 250 grams to Quality of Agricultural Products Department (Wallon Agricultural Research Centre CRA-W, Gembloux, Belgium).

4.3.3 *Fungicide treatment*

Fungicide treatment against leaf diseases, as the leaf rust Roya and Septoria leaf spot (incited by *Puccinia triticina* and *Septoria tritici* respectively) consisted of one application with a concentrated suspension of 12.5% p/v of epoxiconazol, Lovit (Basf España S.A.).

A dose of $1Lha^{-1}$ (the maximum recommended dose) was applied at the phenological phase of flag leaf unfolded, where flag leaf and ligule are just visible. Form this growing stage on leaves are referred in relation to flag leaf. When the crop was in stage 39 of the code of Weber and Bleiholder [26], flag leaf stage: flag leaf fully unrolled, ligule just visible [27-28]. The security time limit of 42 days before harvest was followed.

4.3.4 Chemical analysis

Traditional reference methods were used to compare the quality parameters of both groups of samples. For total content of crude protein the Kjeldahl method was used (Panreac B.O.E 19-7-1977 and 20-7-1977). Moisture was determined by the Panreac air oven method (B.O.E 19-7-1977 and 20-7-1977). Gluten index in order to determine water insoluble protein the Panreac official method was followed (B.O.E 19-7-1977 and 20-7-1977). Finally, total weight of 1000 wheat kernels was performed with Numigral I.

4.3.5 *NIR Spectra*

The spectra were recorded on a *Foss NIRSystems* (model 6500 Foss-NIRSystems, Inc., Silver Spring, Maryland, USA) in reflectance mode, over a wavelength range between 400-2500nm (Visible and NIR region), measured in a 2nm steps.

Intact grain samples were placed in a cuvette of 16.5x3.5cm, with a quartz window (Sample Cell NR-7080) showed in Figure IV-3. In order to obtain useful spectra and avoid outliers two spectra per sample were obtained. The cell was filled with about 20g, were scanned to obtain the spectrum and finally returned to the container and mixed with the remaining sample (approximately 480g) so as to have the maximum variability and information into the spectra. The process was repeated again to obtain the second sample spectrum. To avoid packing variations, only one analyst did sample preparation.

Figure IV-3. Display of wheat intact grains in the cuvette. Wheat with and without treatment (T and O).

Optical density was stored as log (1/R), where R is the reflectance energy recovered by a split detector system with silicon (Si) between 400nm and 1098 nm and a lead sulphide (PbS) between 1100nm and 2500nm.

Computer with the software ISIscan (v2.81; Infrasoft International LLC. Port Matilda now State College, PA, USA) was used for the operation of the spectrometer, and to store and manage optical data.

4.3.6 Hyperspectral imaging

Spectrum data were collected by the near infrared camera, using MatrixNIR Chemical Imaging System (Malvern Instruments, Analytical Hyperspectral NIR system Imaging, Columbia, Mariland, USA) showed in Figure IV-4. It presents an InGas focal-plane array detector (240x320 pixels), working from 900 nm to 1700 nm spectral range by step of 10 nm (which means a total of 85 data points) on reflectance mode.

Fifty random spectra of the total 76800 pixels (240x320 pixels) obtained per image were selected of each image; two images per sample were obtained which means a total of 100 spectra per sample. Mean spectra were computed by averaging the spectra of the kernels within each sample. The software used to spectral selection was Isys^{TM} software (Malvern Instrument Ltd).

The inclusion of a mask helps with the selection of the most representative spectra and avoids the background.

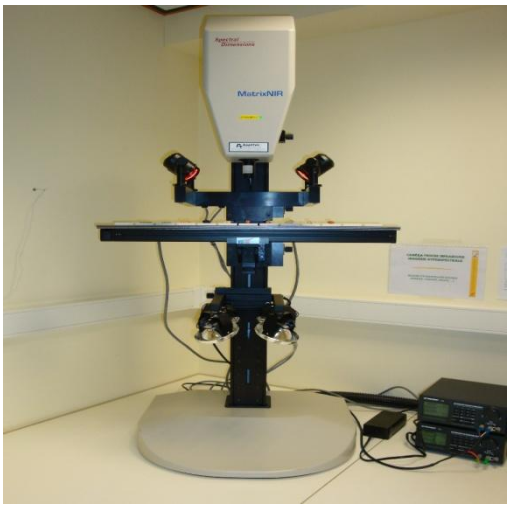


Figure IV-4. MatrixNIR^{TM} Chemical Imaging System instrument

4.4 Statistical analysis and discriminant equations

4.4.1 *Root Mean Squared (RMS)*

Filtering of the subsamples spectra was done by calculating the RMS for this sample presentation form.

The following expressions were applied to calculate RMS values:

$$RMS_j = \sqrt{\frac{\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_j^{\,2})}{n}}$$

$$STD = \sqrt{\frac{\sum_{j=1}^{n}(RMSj)^2}{n-1}}$$

$$STD_{Limit} = 1.036x\sqrt{\frac{\sum_{k=1}^{k=m} STD_k^2}{m}} = 1.036x\sqrt{\overline{STD}^2}$$

Where n is the number of data (absorbance readings), m is the number of samples, Y_{ij} is the absorbance value log (1/R) for sub-spectrum j at wavelength i (λ_i) and $\bar{Y}$ is the absorbance value log (1/R) for the average spectrum of a sample at wavelength i (λ_i).

The STD_{Limit} (Standard Deviation Limit) values were used to obtain $RMS_{Limit.}$ Once the spectra of samples that exceeded the cut-off limit were eliminated, other spectra were obtained on the same sample, obtaining new RMS values. If the new RMS value exceeded the limit again, this sample was marked as not suitable to be included in the calibration set [29].

4.4.2 Calibration and validation sets

NIR: The total sample set used in the study after applying RMS was split into two groups: the first one containing the 80% of the total for construction of the library, which formed the calibration group. Once calibration model were built they were tested one with the remaining 20% for the validating purposes, constituting the second group.

The selection of the sets was carried out in a random way with the option given by the WinISI software: *Functions/ select/ Random samples*. It allows introducing the number of samples to select and split the group.

Hyperspectral: Spectra were imported into the Matlab 7.0 software (version 7.0.0.19920 R14, the Mathworks, Inc.). Prior to calibration development, the data set were split randomly into two groups: training set (80% of the total) used for the calibration model and validation set (20%), which is used as external validation.

4.5 Data analysis

For NIR discriminant analysis based on Principal Component Analysis and Partial Least Square Modified was done with WinISI III software (v1.50e, Infrasoft International LLC).

Hyperspectral imaging was based on Soft Independent Modelling Class Analogy (SIMCA).

4.5.1 Principal Component Analysis (PCA)

Prior to classification models, Principal Components Analysis (PCA), an orthogonal transformation that enables a subspace of R^d to be obtained with a minimum loss of information [30] was performance in both sets (NIR and Hyperspectral). This analysis was explained in section 1.3.3.1.1 of the introduction.

Additionally, in NIR twenty-four Math Pre-treatment which may enhance the accuracy of the final calibration model were applied to spectra. Combinations of derivative (0,0,1; 1,4,4; 2,4,4; 2,10,5), tested on the scatter correction (Standard Normal Variate and Detrend (SNV+DT) and Multiplicative Scatter Correction (MSC)) were used over the spectral range VIS-NIR and NIR.

In hyperspectral imaging Principal component analysis (PCA) was used in order to reduce spectral dimensionality and to visualize the hyperspectral data behaviour of the collective for discrimination. Models were built on raw and with scatter correction: Standard Normal Variate and

Detrend (SNV+D) and Multivariate Scatter Correction (MSC) and first and second derivative.

PCA to reduce dimensionality has been coupled with Mahalanobis distances before carrying out the discriminant analysis [31]. The qualitative analysis between varieties and removal outliers were done with a standardized Mahalanobis (MH) distance. It describes the position in the multidimensional space corresponding to the spectrum of a given sample. Distances between each sample and the population centre greater than 3 are marked as possible spectral outlier [32]. It was decided that sample was eliminated if it appeared as anomalous repeatedly in the different math treatments mentioned.

4.5.2 Partial Least Squares Modified (MPLS)

In NIR the discriminant model was built using Modified Partial Least Squared (MPLS), in winISI software; this assigns to each spectrum a value called a "dummy" variable (or discriminant variable). The new variable obtained, acquired a value between 1 (samples in the group with no fungicide treatment) or 2 (with fungicide treatment group). The discriminant variable limit established for group selection was ≥1.5 [33]. This means that samples with a value lower than 1.5 is included in one group (in this case are O samples) and samples with a high value of 1.5 belong to the other group.

A maximum of twelve PLS terms were selected, if the model selected the 12 PLS terms, the process was repeated with two more terms to avoid overfitting effect. Internal cross validation (with five cross validation groups) was used in order to estimate the final number of PLS terms. Using cross

validation with five groups, on the first pass, the samples of Group 1 are used for the validation, and those remaining four groups are used for the actual calibration. In pass 2, group 2 is used for the actual calibration; in pass 3, group 3, and so on [34]. The math treatments used in both cases were the same as applied in PCA analysis.

The criteria used to select the best models were: Coefficient of determination of calibration (R^2), Standard Error of Cross-Validation (SECV) and samples correctly classified (%).

4.5.3 Soft Independent Modelling Class Analogy (SIMCA)

In hyperspectral NIR imaging SIMCA uses the modelling properties of principal component technique (PCA). So as to search for an optimal classification model, SIMCA algorithm was applied to NIR spectra on raw and after spectral pre-treatment (SNV, MSC, first derivative and second derivative).

The criteria followed for best models selection was % of Correct Classification (%CC).

4.6 Results and Discussion NIR

4.6.1 Prior analysis

When fungal leaf infection is produced early in the season, often prevents the development of the grain. When foliar fungus infection appears, the photosynthesis area decreases which normally means decreases in the

amount of protein, starch and kernels size and weight. However, once the grain has been filled, and subsequently infection occurs, its size will not be affected, even the colour and appearance [35].

Figure IV-5 A represents the raw mean spectra [log 1/R] of samples of fungicide treatment plant and free of it. And IV-5 B after applying second derivative to the mean spectra.

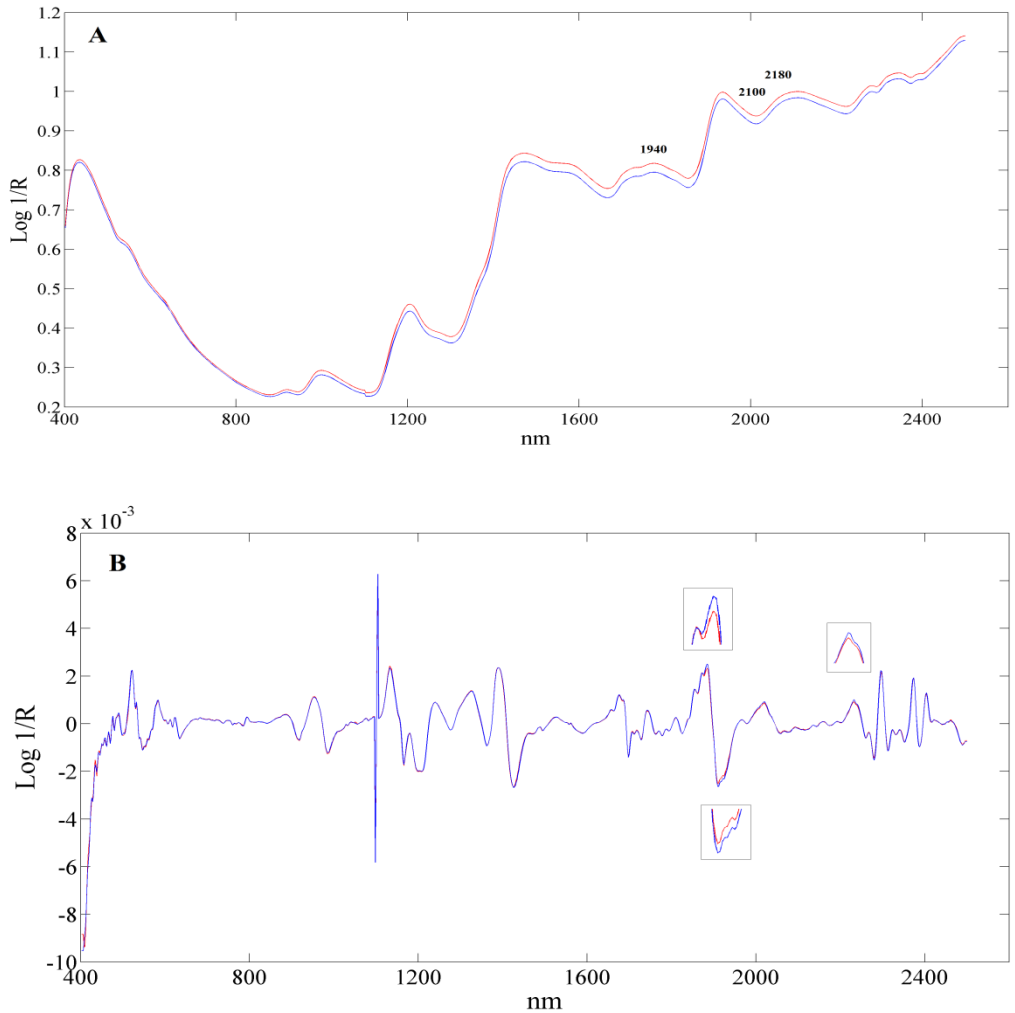


Figure IV-5. A) Average spectra of treated (red) and untreated (blue) samples. B)representation after 2nd derivative.

In the 1200-1318nm, 1440-1880nm and 1920-2498nm wavelength regions there were small differences as much as in both displays, which might be due to different chemical composition in wheat samples. Burns [36] reported that the 1940nm peak is related with moisture in flour, while 2180nm was assigned to protein absorption, 2150nm area has been used for protein [37] and 2100nm for starch. Delwiche [38] indicated that the region 1130-1190nm was a stable region for defining a difference that could be used in classifying normal and scab damaged kernels.

When second derivative is applied differences were reduced, only some peaks mentioned before, such as 1940nm related to moisture, became more evident.

4.6.2 Reference analysis

At first sight there is no evidence of any modification in kernels size or colour by fungus infection. In order to find kernel matrix differences, chemical analysis were performed. There were not significant differences between both groups in moisture or gluten index, being both groups compensated in a similar percentage. On the other hand, protein and 1000 kernel weight showed remarkable grown yield differences.

Figure IV-6 and IV-7 represents the difference between T and O samples. When the difference appears in the positive part means that the measure of T samples is higher than in O samples. Treated samples presented evidence dominance in both parameters; this was expected because when fungal

infected the plant the development of the wheat kernels is reduced and therefore quality parameters.

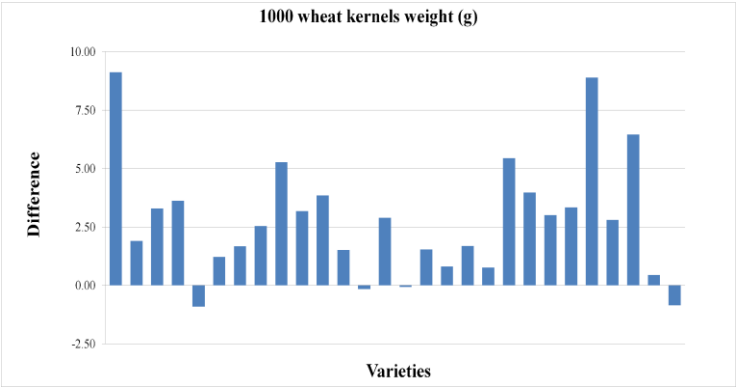


Figure IV-6. Difference between T and O samples in weight of 1000 wheat kernels (grams).

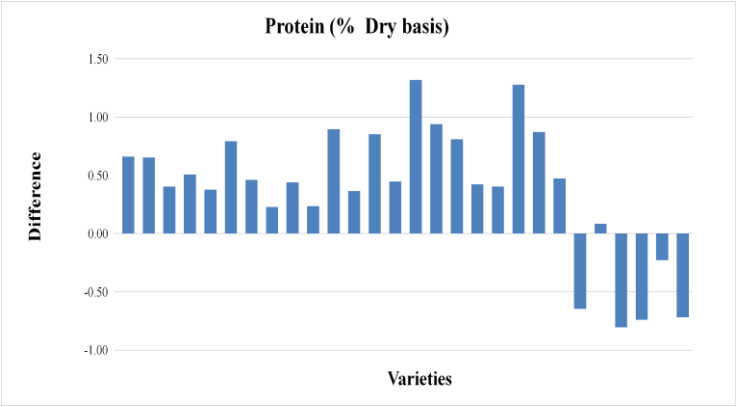


Figure IV-7. Difference between T and O samples in wheat % Protein (Dry basis).

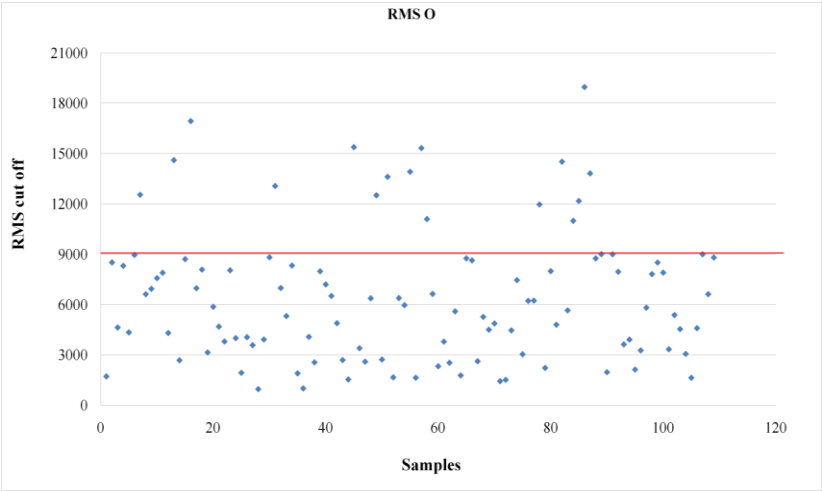
Thus, the parameters protein and thousand grain weight (indicators of the quality of the grain) were higher in treated samples (T), mainly due to the effect that caused the fungicide that prevents the onset of disease.

4.6.3 RMS

So as to perform a filtering process and set the RMS_{Limit}, the values obtained with WinISI software were multiplies by 10^6 this allows a better management of the data.

RMS_{Limit} values were obtained separately for each group. An individual value for each sample was calculated in order to set the maximum value.

It was found that there was a tiny difference between the groups, being establish in 8000 for T samples and 9000 for O samples. Samples which exceeded the RMS cut-off limit were scanned again following the steps detailed in section 2.5.1. Finally a total of 12.5% (14 samples) of T and 14.7% (16 samples) of O were eliminated from whole group (Figure IV-8).



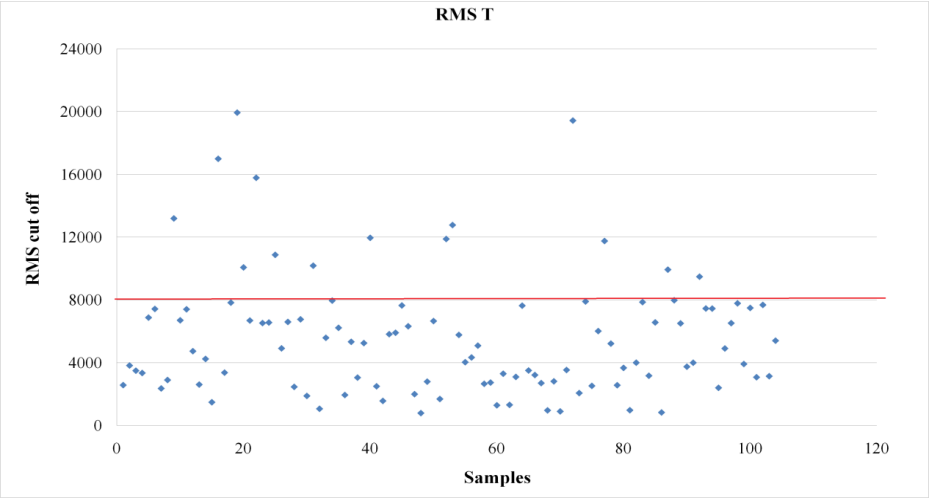


Figure IV-8. Display of the RMS values obtained in each group.

Finally, the remaining spectra were averaged so as to obtain a representative spectrum of each sample.

4.6.4 PCA

To extract initially spectra information and qualitative differences between all samples, data analysis was carried out applying the limit criteria discussed in section 2.4.3. After all the mathematical treatments, only one spectrum was eliminated (T sample).

WinISI program picked 9 factors to cover 99.97% of the explained variance. Before MPLS analysis of the calibration group the accumulative reliabilities of the first 3 PCs were 99.18%, the fourth PC contributed an

additional 0.45% of the total variance, the fifth 0.10% and 0.24% the remaining four.

Using the PCA scores, there was no separation between T and O groups and their corresponding centroids (Figure IV-9).

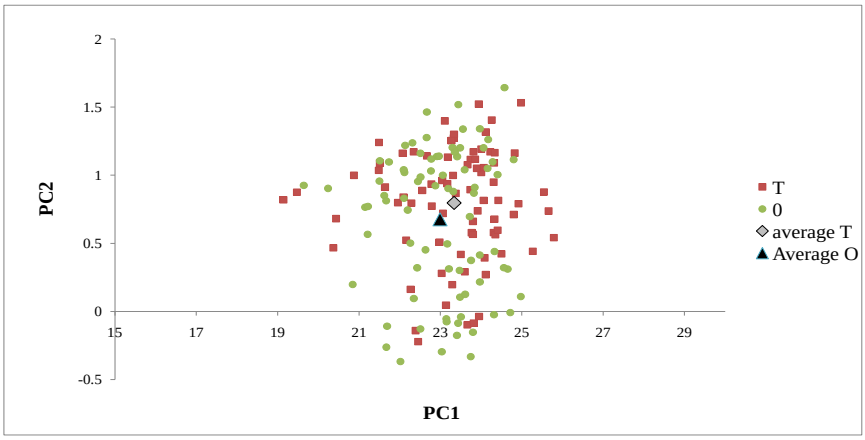


Figure IV-9. Principal Component Analysis: First PC versus Second PC on T (treated) and O (non treated) samples

4.6.5 MPLS

MPLS discriminant models were developed. The dummy variable was set as a referent values for the O and T group. The specification set was 1 for O samples and 2 for T samples.

A maximum of 12 PLS factors and 5 groups of cross validation were used in all the PLS models. After all cross validation passes and with the individual statistical parameter of each group, the number of factors for the smallest error is established. The number of factors required on the spectral MPLS analysis was 7-8 depending on the mathematical treatment applied.

A blind test with 25 samples (12 T and 13 O) was carried out in order to obtain a validation estimate. With appearance of no difference between groups a high percent of correctly classification was still obtained. Table IV-1 A) and B) show the best results obtained on the reflectance mode in NIR and VIS+NIR regions.

Table IV-1. A) Parameter values of the best models developed for VIS+NIR

	Scatter	*Math treatment*	R^2	*SECV*	*Correctly classify*
VIS + NIR 400-2500nm	None	0,0,1	0.50	0.42	0.48
	None	1,4,4	0.58	0.39	0.68
	None	2,4,4	0.76	0.35	0.84
	None	2,10,5	0.71	0.38	0.72
	SNV+DT	0,0,1	0.50	0.42	0.60
	SNV+DT	1,4,4	0.63	0.38	0.52
	SNV+DT	2,4,4	0.74	0.35	0.84
	SNV+DT	2,10,5	0.66	0.37	0.72
	MSC	0,0,1	0.53	0.42	0.60
	MSC	1,4,4	0.73	0.38	0.72
	MSC	2,4,4	0.78	0.34	0.84
	MSC	2,10,5	0.66	0.37	0.68

Table IV-1. B) Parameter values of the best models developed for NIR

	Scatter	*Math treatment*	R^2	*SECV*	*Correctly classify*
NIR 1100-2500nm	None	0,0,1	0.50	0.41	0.56
	None	1,4,4	0.72	0.39	0.76
	None	2,4,4	0.71	0.37	0.80
	None	2,10,5	0.69	0.36	0.76
	SNV+DT	0,0,1	0.55	0.41	0.68
	SNV+DT	1,4,4	0.72	0.38	0.73
	SNV+DT	2,4,4	0.84	0.39	0.76
	SNV+DT	2,10,5	0.69	0.35	0.68
	MSC	0,0,1	0.52	0.40	0.64
	MSC	1,4,4	0.72	0.37	0.72
	MSC	2,4,4	0.72	0.36	0.80
	MSC	2,10,5	0.64	0.37	0.80

The best discriminant model was obtained using the derivative treatment MSC 2,4,4 in VIS and NIR region, which displayed a SECV of 0.34, R^2 of 0.78, with a external validation of 84% of samples correctly classified. Despite those final values of the models for both regions result very similar, statistically better R^2 and lower SECV were obtained when the VIS was included.

Lou [39] and Chandra [40] developed discrimination models to assign categories of wheat kernel damage using a colour machine and hyperspectral image. Their results were similar to those of the Foss NIR system 6500. Figure IV-10 shows a linear representation of the external validation.

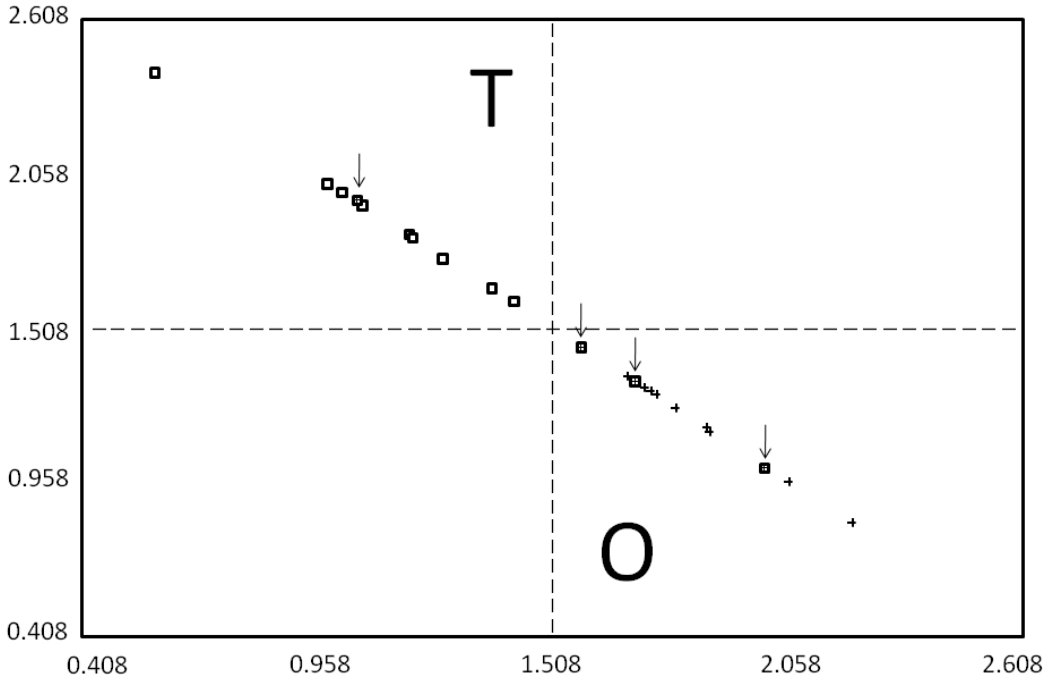


Figure IV-10. Sample misclassification (indicated with an arrow) obtained on the external calibration group.

Quadrant T (T samples) displays one sample belonging to group O; this sample could present a higher resistance to fungus infection which means less modification of the chemical composition. On the other hand, quadrant O presents three samples misclassified, samples coming from treated plants but appearing in the region on untreated samples.

Menniti [41] reported the ineffectiveness of epoxiconazol at controlling some other fungal diseases, which could also be the case in the plants from which these samples come from.

T samples, including as in the group of O, can be due to failure of treatment is not 100%, and therefore has caused disease in some plants.

Conversely, the O samples included in the T group may be because some varieties or growing conditions have allowed disease does not develop in some untreated plants, which are included in the treated group. In addition to these justifications must be added the error of the model itself.

4.7 Results and Discussion Hyperspectral.

4.7.1 *Spectral characteristics*

Once the spectra was obtained, in order to avoid noise interferences it was removal the two extreme sides of the spectra, resulting in a final spectra of 71 nm (960-1670 nm). Figure IV-11 shows the characteristic mean spectra of the 171 durum wheat samples used for calibration, (a) and (b) correspond to the limits established so as to obtain the final spectra without noise.

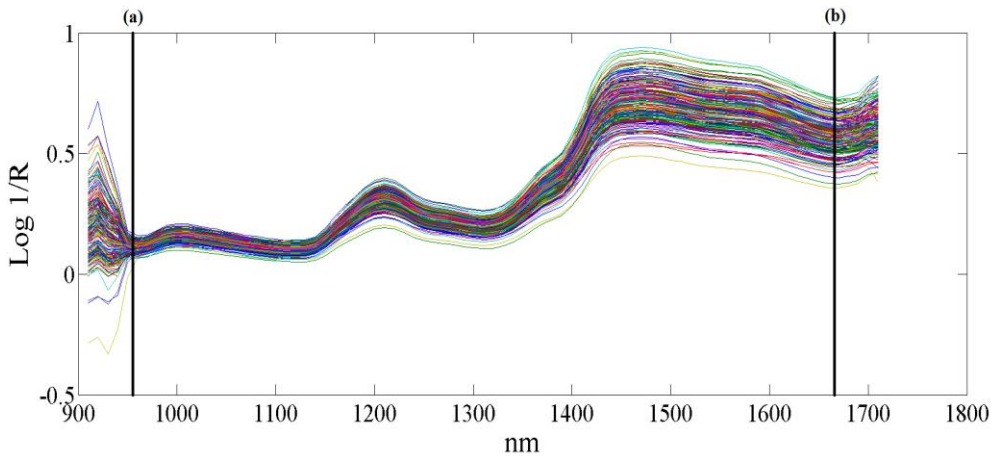


Figure IV-11. Mean Spectral representation of Durum wheat samples

Absorption bands at 1430, 985 and 1210nm are related to starch, where 1430nm is the starch absorbance band [12]; at 960 and 1420nm related to water content and 1470, 1480 and 1500 to protein [43] and [44]

Figure IV-12 (a) shows the original image and sample disposition; the spectral and spatial information were recorded simultaneously providing a NIR spectrum for each pixel in the image of the sample. In order to separate wheat kernels from the image background and avoiding select any interference from background that could influence in the final discrimination a binary mask was created (figure IV-12 (b)) [45].

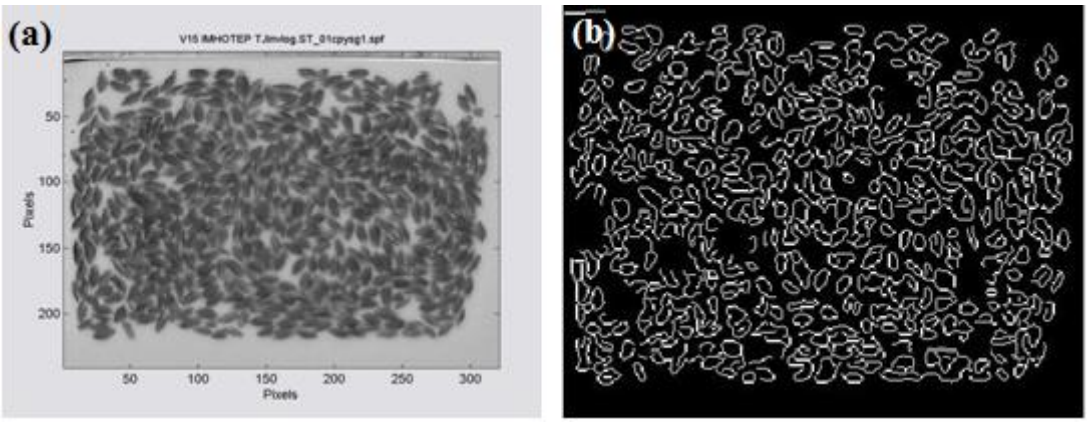


Figure IV-12. (a) Sample presentation and image of the variety Imhotep which comes from Santaella. (b) Mask image of the Imhotep wheat simple

4.7.2 PCA

The principal component analysis was carried out on the calibration set as a visual inspection. Figure IV-13 shows the three dimensional

principal component using the first, second and fourth scores vectors. With the first PC the variance explained of the raw reflectance spectra was 99.95%, the following PC represents only 0.02% of the variance explained.

There is no clear division of the groups. The effect of the infection was barely noticeable in PC3 and was very weak. These score contain information related to kernel classification based on fusarium damage [46].

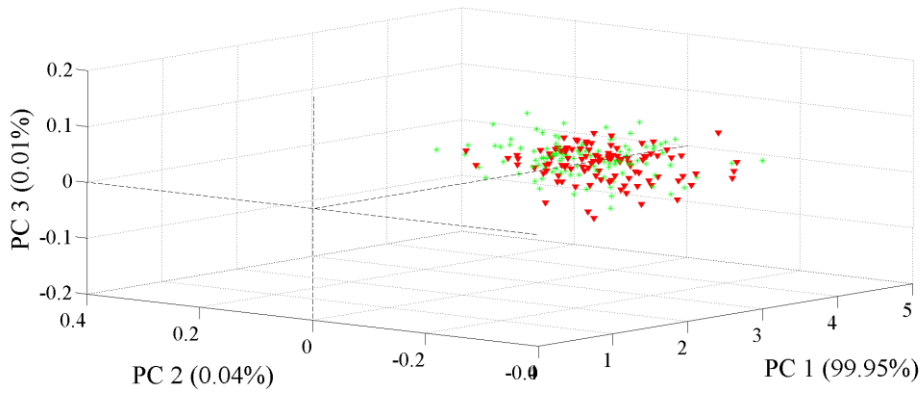


Figure IV-13. Representation of PC1, PC2 and PC3 scores of the calibration group: treated (red) and not treated (green).

The importance of the first principal component in detection of damaged kernels with hyperspectral imaging was reported by Singh 2010 [47] and Shanin, 2011 [48]. Lijuan, 2007, [49] used this technique in order to discriminate transgenic from non-transgenic tomatoes and

Williams 2009 [50], showed well defined clustering corresponding to maize classes.

4.7.3 *SIMCA*

Figure IV-14, IV-15 and IV-16 show the SIMCA classification results obtained by each method to predict samples treated and with no treatment. For every model, a success rate was calculated. An important factor for the classification using SIMCA is the number of PCs to be included in the different models [51].

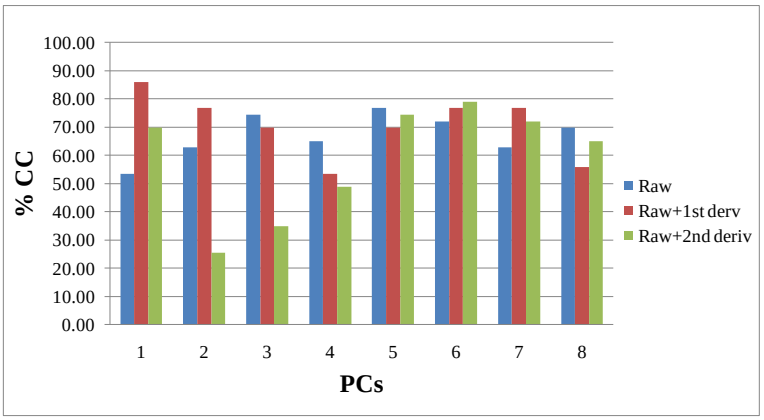


Figure IV-14. Percentage of samples correctly classify (%CC) versus PCs on raw spectra (blue) and after applying 1st derivative (red) and 2nd derivative (green).

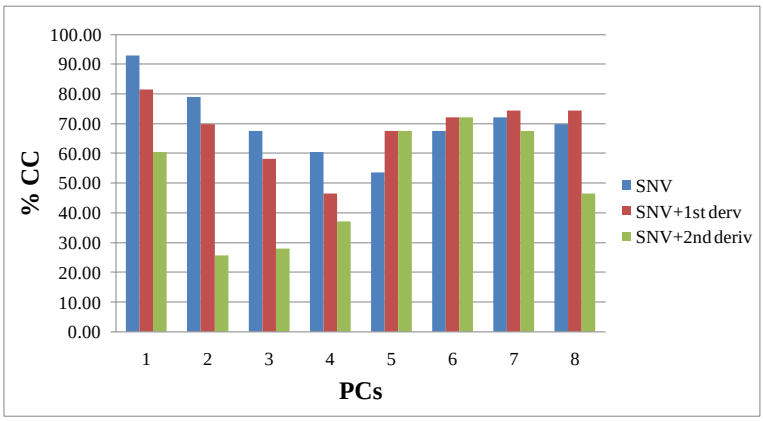


Figure IV-15. Percentage of samples correctly classify (%CC) versus PCs after SNV (blue), SNV+ 1st derivative (red) and SNV+2nd derivative (green).

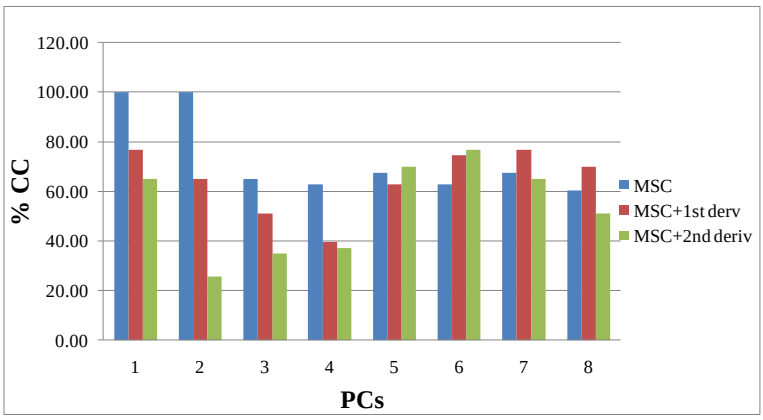


Figure IV-16. Percentage of samples correctly classify (%CC) versus PCs after MSC (blue), MSC+ 1st derivative (red) and MSC+2nd derivative (green).

The success rate was calculated. The accuracy of these mathematical models showed that SIMCA can successfully discriminate between samples that have received treatment and those with no treatment. Preprocessing the spectra did influence in the results. However, there is a clear reduction of the

model capacity of predicting when second derivative is applied to the spectral, caused by the lost of information (particle size, packaging…).

The best models obtained with a 100% of correctly classification correspond to MSC with no derivative and using the first two PCs, which apparently have the majority of the information to discriminate the wheat samples. Zahn, 2010 [52] reported that the first few PCs were found to be sensitive to the spectra of the healthy and stressed plants. He worked with fungal infection levels in rice and hyperspectral obtained the same result applying Learning Vector Quantization (LVQ) with PCA and second derivative. However, Esteve 2012 [53] reported that after applying scatter correction to the spectra of sound and heat-damaged kernels the misclassification increased.

Bibliography

[1] (www.fao.org, 2010).

[2] Anuario de Estadística Ministerio de Medio Ambiente y Medio Rural y Marino. Marm, 2009.

[3] F.M. Dupont and S.B. Altenbach. Molecular and biochemical impacts of environmental factors on wheat grain development and proteins synthesis. Journal of Cereal Science (2003) 38 133–146.

[4] N. K. Fageria, V. C. Baligar and C. A. Jones. Growth and Mineral Nutrition of Field Crops, Third Edition (2011). CRC Press.

[5] Z. Eyal, A. L. Scharen, J. M. Prescott and M. van Ginkel. Enfermedades del trigo causadas por Septoria: Conceptos y métodos relacionados con el manejo de estas enfermedades. México, D.F; CIMMYT (1987). 45p.

[6] M. D. Bolton, J. A. Kolmer and D. F. Garvin. 2008. Wheat leat rust caused by Puccinia tritinica. Molecular Plant Pathology (2008) 9(5), 563-575.

[7] L. López Bellido. Cultivos herbáceos. Vol. I. Cereales. Ed Mundi-Prensa, Madrid, Spain, (1991) 539 pp.

[8] www.alimentacion.org.ar.

[9] X.M. Chen. Epidemiology and control of stripe rust [*Puccinia striiformis* f. sp. *tritici*] on wheat. Can. J. Plant Pathol. (2005) 27: 314–337.

[10] R. M. Balabin and E.I. Lomakina. Support Vector Machine regression (SVR/LS-SVM)- an alternative to Neuronal Networks (ANN) for analytical chemistry? Comparison of nonlinear methods on Near Infrared (NIR) Spectroscopy data. Analyst 136, (2011) 1703.

[11] R. M. Balabin, R. Z. Safieva. 2008. Motor oil classification by base stock and viscosity based on near infrared (NIR) spectroscopy data. Fuel 87(2008) 2745.

[12] R. M. Balabin, R. Z. Safieva, E.I. Lomakina. 2010.Gasoline classification ising near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. Anal. Chim. Acta 671, (2010) 27.

[13] R. M. Balabin, R. Z. Safieva and E.I. Lomakina. Near-Infrared (NIR) Spectroscopy for motor oil classification: From discriminant analysis to support vector machines. Microchem. J. 98, (2011) 121.

[14] A. Alishahi, H. Farahmand, N. Prieto and D. Cozzolino.2010. Identification of transgenic foods using NIR spectroscopy: A review. Spectrochimica Acta Part A 75 1-7.

[15] G. Downey. Authentification of food and food ingredients by near infrared spectroscopy. J. Near Infrared Spectrosc. 4, 47-61 (1996).

[16] M. Kubik. Physical Techniques in the Study of Art, Archaeology and Cultural Heritage Volume 2. Chapter 5, Hyperspectral Imaging: A New Technique for the Non-Invasive Study of Artworks. Elsevier (2007).

[17] G. ElMasrya and D. W. Sun. Chapter 1: Principles of Hyperspectral Imaging Technology. Hyperspectral Imaging for Food Quality Analysis and Control. Elsevier (2010).

[18] G. ElMasrya, N. Wanga, C. Vigneault, J. Qiao and A. ElSayed. Early detection of apple bruises on different background colors using hyperspectral imaging. LWT (2008) 41 337–345.

[19] G. ElMasry, N. Wang, A. ElSayed and M. Ngadi. Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry. Journal of Food Engineering 81 (2007) 98–107.

[20] A. A. Gowen, M. Taghizadeh and C. P. O'Donnell. Chapter 13: Using Hyperspectral Imaging for Quality Evaluation of Mushrooms. Hyperspectral Imaging for Food Quality Analysis and Control. Elsevier Inc. (2010).

[21] M. Taghizadeh, A. A. Gowen and C. P. O'Donnell. The potential of visible-near infrared hyperspectral imaging to discriminate between casing soil, enzymatic browning and undamaged tissue on mushroom (Agaricus bisporus) surfaces. Computers and Electronics in Agriculture 77 (2011) 74–80.

[22] P. Rajkumar, N. Wang, G. EImasry, G.S.V. Raghavan and Y. Gariepy. Studies on banana fruit quality and maturity stages using hyperspectral imaging. Journal of Food Engineering 108 (2012) 194–200.

[23] J.A. Fernández Pierna, V. Baeten and P. Dardenne. Screening of compound feeds using NIR hyperspectral data. Chemometrics and Intelligent Laboratory Systems 84 (2006) 114–118.

[24] H. Yao and D. Lewis. Chapter 2: Spectral Preprocessing and Calibration Techniques. Hyperspectral Imaging for Food Quality Analysis and Control. Elsevier (2010).

[25] K. Bhuvaneswari, P. G. Fields, N. D.G. White, A. K. Sarkar, C. B. Singh and D. S. Jayas. Image analysis for detecting insect fragments in semolina. Journal of Stored Products Research 47 (2011) 20-24.

[26] E. Weber and H. Bleiholder. Erfäuterungen zu den BBCH-Decimal-Codes für die Entwicklungsstadien von Mais, Raps, Faba-Bohne, Sonnenblume und Erbsema Abbildungen. Gesunde Pflanzen (1990) 42, 308-321.

[27] U. Meier. Estadios de las plantas mono- y dicotiledóneas. BBCH Monografía, Centro Federal de Investigaciones Biológicas para Agricultura y Silvicultura (2001).

[28] M. Cátedra Cerón and I. Solís Martel. Effect of a fungicide treatment on yield and quality parameters of new varieties of durum wheat (Triticum turgidum L. ssp. Durum) and bread wheat (triticum aestivum L.) in western Andalusia. Spanish Journal of Agricultural Research (2003) 1 (3), 19-26.

[29] A. J. Gaitán-Jurado, M. García-Molina, F. Peña-Rodriguez and V. Ortiz-Somovilla. Near Infrared applications in the quality control of seed cotton. J. Near Infrared Spectrosc. (2008) 16,421-429.

[30] Y. Langeron, M. Doussot, D.J. Hewson and J. Duchêne. Classifying NIR spectra of textile products with kernel methods. Engineering Applications of Artificial Intelligence (2007) volume 20, issue 3.

[31] B. G. Osborne, T. Fearn, and P. H. Hindle. Near infrared calibration II. Ch.7 In: Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, 2nd ed. Longman Scientific & Technical, UK (1993) 121-144 pp.

[32] J. S. Shenk and M. O. Westerhaus. Population structuring of near infrared spectra and modified partial least square regression. *Crop Science* (1991) 31, 1548-1555

[33] G. Downey. Discrimination et authentification des aliments et des ingredients alimentaires par spectroscopie dans l'infrarouge proche et moyen. In: La spectroscopie infrarouge et ses applications analytiques. Bertrand, D. and Dufour, E. (eds). Editions TEC & DOC, Paris, France (2000) pp. 397-422.

[34] www.zeiss.com.

[35] A. Okorski, J. Olszewski, A. Pszczółkowska and T. Kulik. Effect of fungal Infection and the Application of the Biological Agent EM 1™ on the Rate of Photosynthesis and Transpiration in Pea (Pisum Sativum L.) Leaves. Pol. J. Natur. Sc. (2008) Vol 23(1): 35-47, Y.

[36] A. Burns and E. W. Ciurczak. Handbook of Near-Infrared Analysis Third Edition. CRC Press (2008).

[37] A. Ruano-Ramos, A. García-Ciudad and B. García-Criado. Determination of nitrogen and ash contents in total herbage and botanical components of grassland systems with near infrared spectroscopy. Journal of the Science of Food and Agriculture (1999) 79:137-143.

[38] S. R. Delwiche and G. A. Hareland. Detection of Scab-Damaged Hard Red Spring Wheat Kernels by Near-Infrared Reflectance. Cereal Chem. (2004) 81(5):643–649.

[39] X. Lou, D.S. Jayas and S. J. Symons. Identification of Damage Kernels in Wheat using a Colour Machine Vision System. Journal of Cereal Science 30(1999) 49-59.

[40] C. B. Singh, D. S. Jayas, J. Paliwal and N. D. G. White. Detection of midge-damaged wheat kernels using short-wave near-infrared hyperspectral and digital colour imaging. Biosystem Engineering 105 (2010) 380-387.

[41] A. M. Menntiti, D. Pancaldi, M. Maccaferri and L. Casalini. Effect of fungicides on Fusarium head blight and deoxynivalenol content in durum wheat grain. European Journal of Plant Pathology 109: (2003) 109-115.

[42] P. C. Williams. Near-infrared spectra. Near Infrared Technology in the Agricultural and Food Industries. P. Williams and K. Norris. eds. AACC International: St. Paul, MN. (2001) Pages 239-282.

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

[44] I. Murray and P. C. Williams. Chemical principles of nearinfrared technology. In: Near-infrared Technology in the Agricultural and Food Industries (Williams P C; Norris K H eds). American Association of Cereal Chemists Inc., St. Paul, Minnesota (1987) pp. 17–34.

[45] J. Li, X. Rao and Y. Ying. Detection of common defects on oranges using hyperspectral reflectance imaging. Computers and Electronics in Agriculture 78 (2011) 38–48.

[46] M. A. Shahin and S. J. Symons. Detection of Fusarium damaged kernels in Canada Western Red Spring wheat using visible/near-infrared hyperspectral imaging and principal component analysis. Computers and Electronics in Agriculture 75 (2011) 107–112.

[47] C. B. Singh, D. S. Jayas, J. Paliwal and N. D.G. White. Detection of midge-damaged wheat kernels using short-wave near-infrared hyperspectral and digital colour imaging. Biosystems engineering 105 (2010) 380-387.

[48] M. A. Shahin and S. J. Symons. Detection of Fusarium damaged kernels in Canada Western Red Spring wheat using visible/near-infrared hyperspectral imaging and principal component analysis. Computers and Electronics in Agriculture 75 (2011) 107–112.

[49] L. Xie, Y. Ying, T. Ying, H. Yu and X. Fu. Discrimination of transgenic tomatoes based on visible/near-infrared spectra. Analytica Chimica Acta 584 (2007) 379–384.

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

[51] J. A. Fernández Pierna, P. Volery, R. Besson, V. Baeten and P. Dardenne. Classification of Modified Starches by Fourier Transform Infrared Spectroscopy Using Support Vector Machines. J. Agric. Food Chem. (2005**)** 53, 6581-6585.

[52] L. Zhan-Yu, W. Hong-Feng and H. Jing-Feng. Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis. Computers and Electronics in Agriculture 72 (2010) 99–106.

[53] L. Esteve Agelet, D. D. Ellis, S. Duvick, S. Goggi, Charles R. Hurburgh and Candice A. Gardner. Feasibility of near infrared spectroscopy for analyzing corn kernel damage and viability of soybean and corn kernels. Journal of Cereal Science 55 (2012) 160-165.

Chapter V

Application of NIRS in authentication of bread wheat varieties from southern Spain

5.1 Introduction

Cereals constitute a significant food resource for the world's population [1]. Wheat is the second largest grain crop of Spain. During the latest decade an average of 506,745 hectares in Andalusia were dedicated to wheat cultivation, of which 126,981 ha were used for bread wheat and 378,006 ha to durum wheat. Because of it geographical and climatic diversity, Andalusia can produces many varieties of wheat.

The application of the European agrarian politics and the regulation on domestic wheat quality has made farmers improve their competitiveness through standardization [2]. The use quality is influenced 60% by genetics and 40% by management practice.

Werner, 2006 [3] told that "Knowledge of varieties was still low". Food quality of wheat is obviously affected by many factors, such as variety and growing conditions [4]. The quality of wheat grain for milling industry is

determined by the protein quantity, quality and degree of starch damage and by alpha amylase content. However, this concept of quality is complex because not only protein quantity is important but also protein type.

The wheat delivery to the market is an essential point in obtaining homogeneous flours. Misclassification may result in a decreased quality and uniformity of the final product [5]. The problem found during this process is the determination of the physiochemical properties (such falling number, elasticity, extensibility) of bread wheat that determine the quality of the final flour. These methods include polyacrylamide gel electrophoresis [6], High-Performance Liquid Chromatography (HPLC) [7], Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) [8], all are time consuming, tedious and not suited for real time the variety identification.

NIR has been used for many years in the milling industry. The use of this technology has allowed rapid and accurate measurement of parameter such protein [9]; [10]; [11], hardness [12]; [13] and moisture [14]. However, there are not completely descriptive of wheat use suitably.

5.2 Objective

The goal of the present study is to use NIR to discriminate between bread wheat samples considering chemical properties and variety.

5.3 Materials and Methods

5.3.1 *Wheat samples*

The wheat samples came from the Andalusia Network of Agrarian Experimentation (RAEA), over the crop seasons 2004-2008, in different areas of Andalusia.

Five wheat soft varieties "Cartaya"=37, "Odiel"=17, "Gazul"=38, "Yecora"=29, "Galeón"=33 were used in this work, giving 154 samples. Yecora, Galeón, Cartaya and Gazul were the more outstanding varieties of bread wheat along 2005-2008 crop seasons.

Quality analysis of bread wheat varieties are:

Protein: crucial in relation of breadmaking quality. It forms and stabilizes the foam structure of bread.

Tenacity (T) and the Extensibility (E) relation T/E: Other quality parameters used by the flour industry, it indicates the balance of the flour and express what type of work is more suitable to each one.

$T/E<0.5$ extensible flour

$0.5<T/E<0.8$ balance flour

$T/E>0.8$ robust

Baking strength (W): Expresses the baking strength and indicates the work needed to break a sheet of dough pushed by air. W is represented by the curve area of the alveogram.

W>300 strong flour

150<W<300 medium strength flour

80<W<150 normal flour

W<80 non bread flour

Zeleny sedimentation value: gluten quality measurement, higher Zeleny value better gluten quality. Acceptable bread wheat should have a minimum of 22ml.

Falling number: measure the alpha amylase activity. A value below 180 indicates high amylase activity, resulting in non bread flour. Correct index values are between 250 and 300 seconds.

Specific weight: it measures a ratio between weight and volume of a given sample. This parameter is a good estimate of the physical quality of the grain and milling performance.

5.3.2 *NIR spectra*

Bread wheat samples were provided by the Red Andaluza de Experimentación Agraria (RAEA), managed by the Instituto de Investigación y Formación Agraria y Pesquera (IFAPA). Samples originated from four trial sites located in Jerez (Cadiz), Granada, Tomejil and Carmona (Seville) and Cañete de las torres (Cordoba), each having different agroclimatic conditions.

The procedure to collect the samples was the same as described in Chapter 3. The IFAPA Center Alameda del Obispo (Córdoba) received the bread wheat seeds in bags containing approximately 500g. They were transferred to small plastic containers with a capacity of approximately 250 g.

To obtain the spectral information approximately 22 g of each sample were used. The NIRS instrument was a Foss NIRSystems 6500 (Foss-NIRSystem, Inc., Silver Spring, Maryland, USA) working in reflectance, two detectors Silicon: (Si) 400-1100nm and 1100- 2500nm Lead Sulphide (PbS). Wavelength increment was 2nm.

Samples were set in a cuvette with 16.5 x 3.5 cm. A quartz window allows the incident light pass through it. Figure V-1 A and B display the cuvette used for spectral acquisition and a sample presentation.

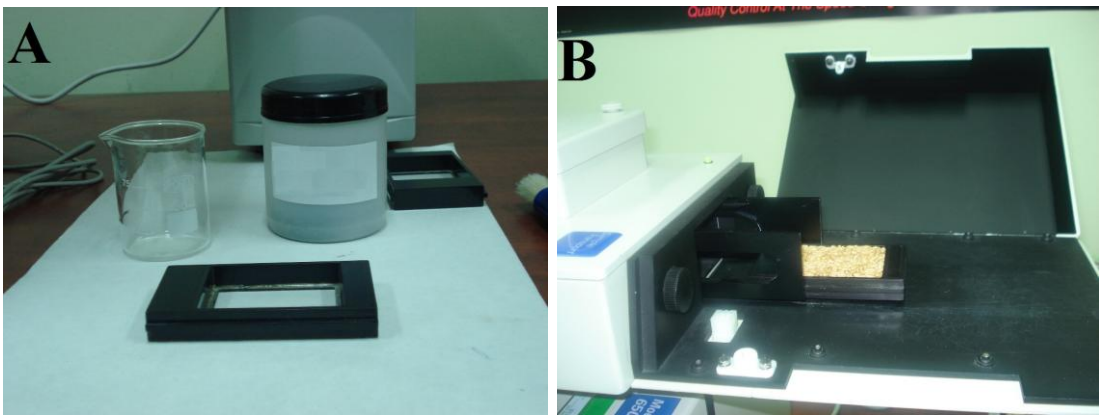


Figure V-1. A) cuvette used for placing the sample and taking the spectra. B) sample presentation before being scanned.

5.3.3 Calibration and validation groups

This part follows the same pattern as the one found in chapter 5 section 3.4.2. The collective was randomly divided in two groups, the first one for calibration the 75% of the total set, and the second for validation the remaining 25%.

Final dataset was set into 116 and 38 for calibration and validation, respectively.

5.3.4 *Statistical and discriminant analysis*

Multivariant and statistical analysis was carried out in this work with WinISI software v. 1.50e (Infrasoft International, Port Matilda, Pennsylvania, USA).

5.3.41 *Principal Component Analysis (PCA)*

As a previous step to the development of qualitative models, PCA was applied (described in section 1.3.3.1.1 of chapter 1). The aim of this analysis was to identify samples that could be as spectral anomalous (outliers). In this case the outlier detector Mahalanobis distance (MD) was used. Each sample which distance was greater than 3 [15], [16] was marked as outlier and not included into the calibration set.

The same scatter correction and spectral pre-treatment as in the following section (Modified Partial Squares) were used.

5.3.42 *Modified Partial Squares equation (MPLS)*

The equations were obtained using the regression method MPLS (explained in section 3.5.2). Ten blocks cross validation was used to optimize the number of latent variables.

The models were evaluated with and without pre-treatments of the spectra data; combining raw spectra with derivative treatments of 0,0,1, 1,4,4 and 2,15,8 [5] and the same combination with Standard Normal Variate and Detrend (SNV+DT) and Weighted Multiplicative Scatter Correction (WMSC) so as to eliminate the scatter interferences. The following spectral regions were used: Visible +Near Infrared (VIS+NIR 400-2500nm) and Near Infrared (NIR 1100-2500nm) (Table V -1).

Table V-1. Pre-treatment combination for developing the models

	Spectral Range	*Scatter Correction*	*Derivative*
Vis + NIR	(400-2498nm)	None	0,0,1
		SNV+DT	1,4,4
NIR	(1100-2498nm)	WMSC	2,15,8

Best discriminant models were selected according to the lower error of cross validation (SECV) in calibration and percentage (%) of less samples misclassified in validation.

5.4 Results and Discussion

5.4.1 *Wheat characterization*

Table IV-2 displays mean quality results and it Standard deviation (SD) of bread wheat samples from 2004-2008 crop seasons.

Table V-2. Quality parameters of soft wheat samples

		Protein	T/E	(W)	Zeleny sedimentation value	Falling number	Specific weight Kg/Hl
Cartaya	Mean	12.48	1.13	181.74	29.62	332.15	80.40
	SD	1.14	0.54	48.59	3.88	114.12	3.67
Gazul	Mean	13.96	1.21	337.10	40.56	393.12	82.00
	SD	1.28	0.83	102.06	4.25	18.27	2.71
Galeón	Mean	13.81	1.09	232.20	39.42	410.44	80.92
	SD	1.39	0.70	61.53	8.03	18.64	2.67
Yecora	Mean	13.98	1.12	315.00	40.62	393.56	81.49
	SD	1.41	0.74	99.87	5.65	21.71	2.82
Odiel	Mean	11.69	0.88	66.14	26.14	356.43	81.60
	SD	1.10	0.92	25.61	8.91	20.85	1.24

Cartaya and Odiel have the lowest values in W and Falling number. Odiel with 66.14 and Cartaya with 181.74 in W, result of non bread and medium strength flour, respectively. Gazul, Galeón and Yecora present slight variations in quality parameters. T/E relation rise above 0.8 in all varieties as consequence of robust flour.

5.4.2 *Spectral characterization*

The resulting representative spectra of all the samples are showed in figure IV-2.

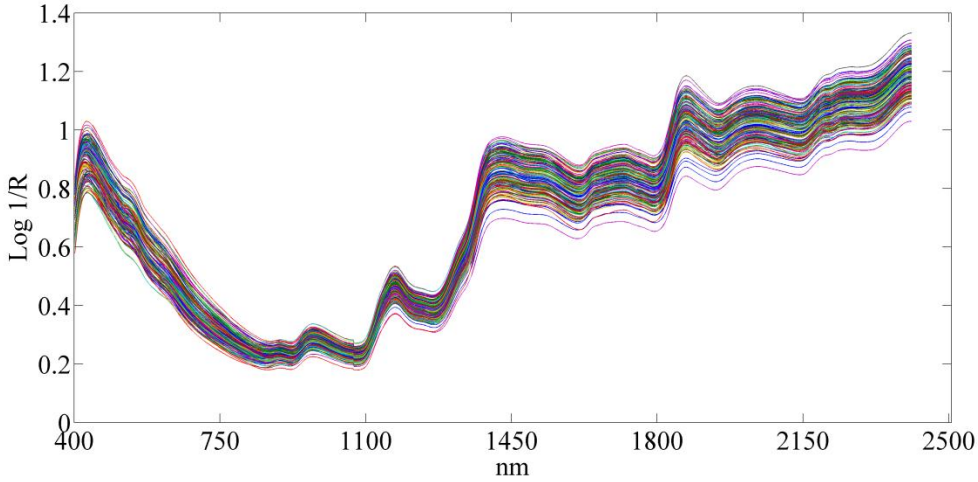
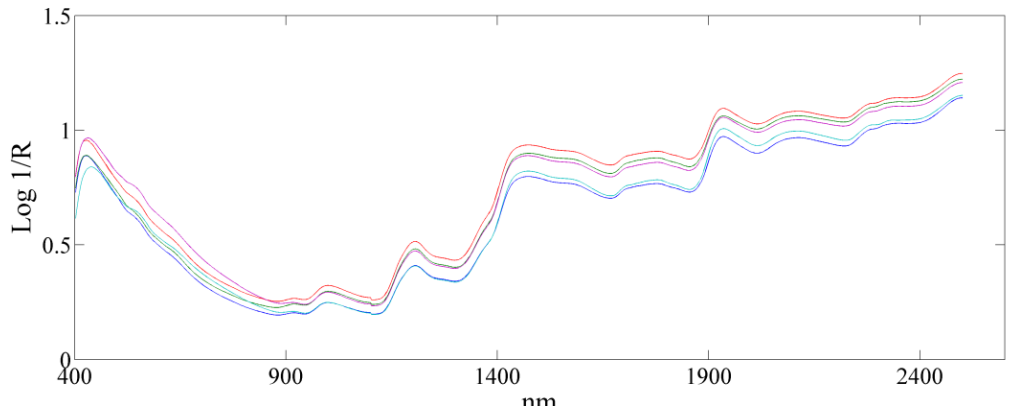


Figure V-2. Characteristic raw spectra of bread wheat samples.

Spectra characteristics are influenced by size, shape and chemical composition [13]. Figure V-3 A) displays the average spectra of the five varieties of wheat and B) after SNV+DT and second derivative were applied.

(A)



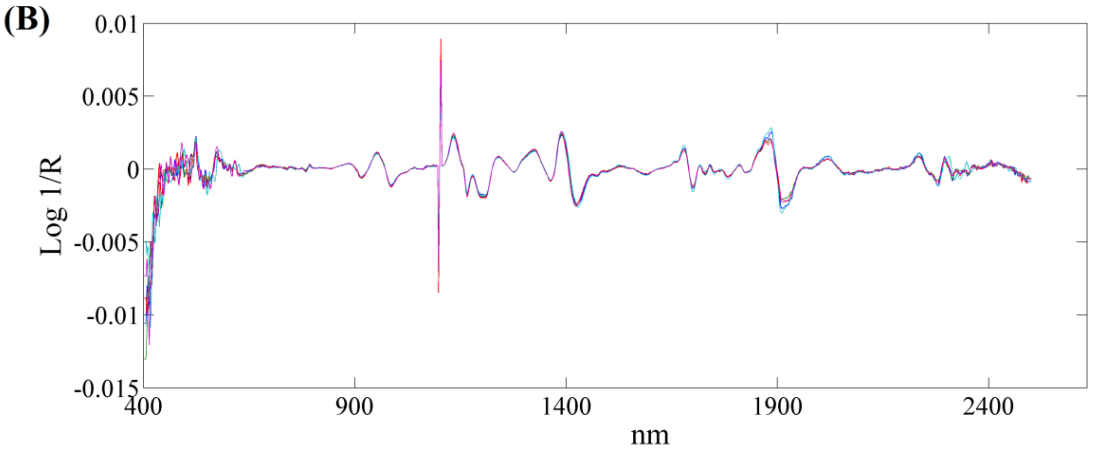


Figure V-3. (A)Representation of mean spectrum corresponding to the five varieties of bread wheat samples. (B) after applying spectral pre-treatment 2º derivative and SNV+DT.

When no pre-treatment was applied the spectra followed the same pattern despite differences on reflectance light caused by multiplicative and additive effects. Peaks at 1480 correspond to protein and starch [17], [18]; at 1420 and 1900nm are close to absorption peaks caused by water [13] and 1390 to oil [19]. Second derivative eliminated effects such as particle size or sample packaging. The major spectral differences were in the 1900-2000 nm corresponding to water band region [17].

5.4.3 PCA

Figure V-4 shows the representation of the three first principal components after applying SNV+DT and 2nd derivative. With only the first two principal components the 95.76% of the variance was explained.

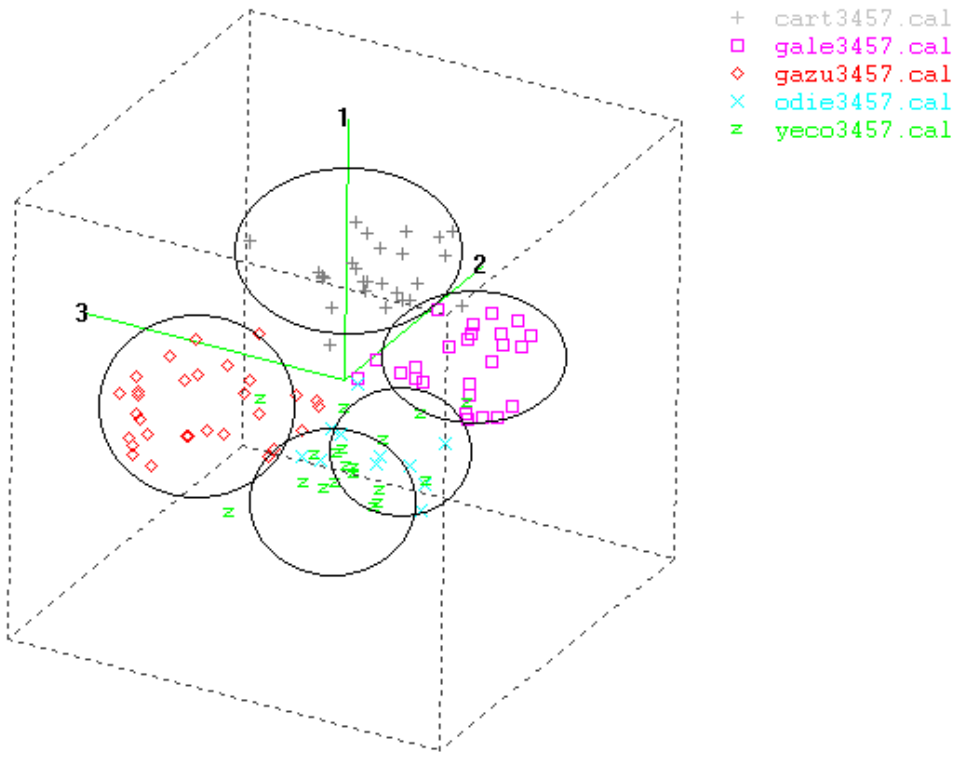


Figure V-4. PC1, PC2 and PC3 analysis of the five wheat samples: Cartaya (grey), Gazul (red), Galeón (pink), Yecora (green) and Odiel (blue).

It can be observed five clusters related to the varieties. Even coming from different parts of Andalusia with their corresponding agroclimatics characteristics, there is an obvious grouping caused by varietal features resulting in a detectable effect on the NIR spectra. Varieties were grouped even though each variety was grown in all regions.

Three samples were eliminated from the data set as consequence of the Mahalanobis distance (MD) which values were greater than 3. The final distribution of the samples according to the MD values are represented in figure V-5.

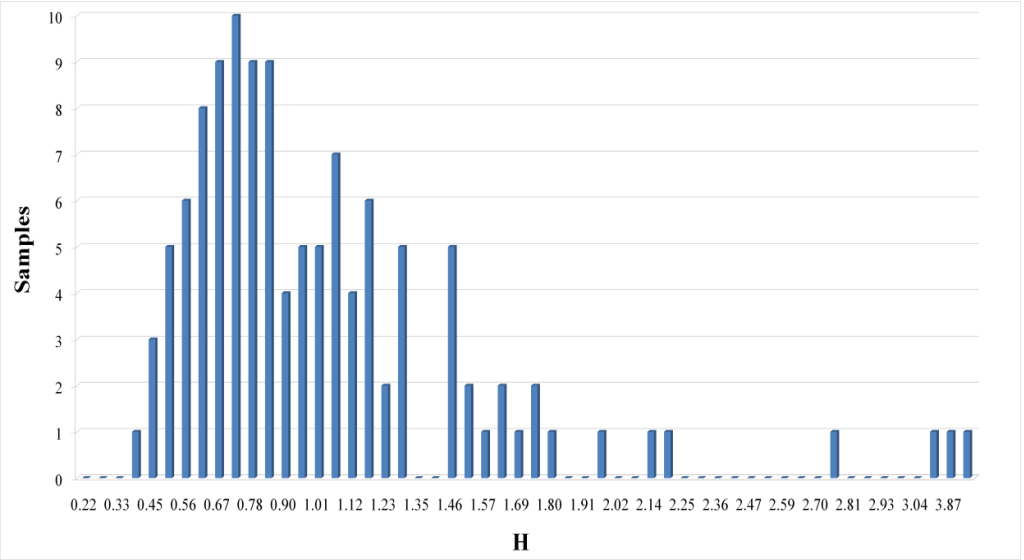


Figure V-5. Mahalanobis distances of the group of samples.

The histogram shows the distribution of the Mahalanobis (MD) distance for the bread wheat varieties. Each bar represents the number of samples within their situation over the space and the group follows a Gaussian distribution where most of the samples are placed in the centre of the histogram.

5.4.4 Discriminant equations and external validation

Models developed follow the combination of the pre-treatment described in section 4.2. There were carried out 5 cross validation segments with a maximum of 12 PLS terms. Table V-3 shows the results obtained

with the models over the VIS+NIR region on raw (none) and after applying SNV+DT and WMSC; table V-4 only with the NIR region.

Table V-3. Results of the discriminant analysis on VIS+NIR

	Scatter correction	Derivative	PLS factors	Error of Classification (%)	SECV	r^2
VIS + NIR	None	0,0,1	10	13,51	0,26	0,53
	None	1,4,4	10	5,40	0,23	0,64
	None	2,15,8	9	4,50	0,23	0,64
	SNV+DT	0,0,1	10	9,00	0,26	0,55
	SNV+DT	1,4,4	9	4,50	0,23	0,64
	SNV+DT	2,15,8	9	4,50	0,23	0,66
	WMSC	0,0,1	10	8,10	0,25	0,58
	WMSC	1,4,4	9	4,50	0,23	0,64
	WMSC	2,15,8	9	4,50	0,23	0,66

Table V-4. Results of the discriminant analysis on NIR

	Scatter correction	Derivative	PLS factors	Error of Classification (%)	SECV	r^2
NIR	None	0,0,1	10	36,94	0,34	0,24
	None	1,4,4	10	18,01	0,28	0,49
	None	2,15,8	10	18,01	0,29	0,46
	SNV+DT	0,0,1	10	32,43	0,32	0,34
	SNV+DT	1,4,4	10	16,22	0,28	0,50
	SNV+DT	2,15,8	10	13,51	0,28	0,51
	WMSC	0,0,1	10	33,33	0,34	0,28
	WMSC	1,4,4	10	13,51	0,28	0,50
	WMSC	2,15,8	10	12,61	0,28	0,49

Results obtained for the 18 analysis including both regions showed that equations developed over the VIS+NIR region had lower error of classification that NIR one, accounting in some models with one additional pls factor. Features related to visible region have demonstrated high discrimination capability in previous statistical classification models [20]. Pigmentation was sufficiently different in the types of kernels over the 400-780nm visible range to be detected by the silicon detector.

It can be seen that when applying either SNV+DT or MSC the accuracy of the models improve considerably. Derivatives were necessary to reduce error. Over the 400-2500nm the error of classification when no derivative is used is two times higher than when it was used. In NIR region the impact of derivatives is even larger (three times than when it is applied).

In the two models based in VIS+NIR region with SNV+DT+2º derivative and WMSC+2º derivative the 95,5% of the variance is explained. Classification matrix is showed in figure V-6.

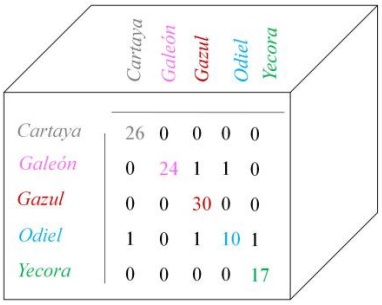


Figure V-6. Classification matrix of the discriminant model WMSC + 2º derv.

Cartaya, Gazul and Yecora are perfectly differentiated from the rest of the groups with no erroneous classification. Galeón identified as Gazul and Odiel. Three samples of Odiel were situated in the groups Cartaya, Gazul and Yecora.

Table V-5 represents the confusion matrix of the external validation with the best model used.

Table V-5. Confusion matrix of external validation

Varieties	Corretly classify	Erroneously classify
Cartaya	9	1
Gazul	6	0
Yecora	10	0
Odiel	4	1
Galeón	7	0

The results displayed in the table agreed with the original PC analysis. Samples erroneously classified only represent 5.26%.

Bibliography

[1] A. Aydin, P. Paulsen and F. J.M. Smulders. The physico-chemical and microbiological properties of wheat flour in Thrace. Turk J Agric For 33 (2009) 445-454

[2] M. Guerrero. Tepid implementation of the Domestic Wheat Quality Standardization Regulation. USDA Foreign Agricultural Service, 2011.

[3] W. Vogt-Kaute. Baking quality of variety mixtures of organic wheat and acknowledgement by farmers and millers in Germany. Proceedings of the cost susvar workshop on Cereal crop diversity: implications for production and products. 2006.

[4] K. Miezan, E.G. Heyne and K. F. Finney. Genetic and environmental effects on the grain protein content in wheat. Crop Science (1977) 17:591-593.

[5] C. Miralbés. Discrimination of European wheat varieties using near infrared reflectance spectroscopy. Food Chemistry 106 (2008) 386–389.

[6] K. Khan, C. E. McDonald, and O. J. Banasik. Polyacrylamide Gel Electrophoresis of Gliadin Proteins for Wheat Variety Identification-Procedural Modifications and Observations. Cereal Chem. (1983) 60(2):178:181.

[7] T. Dachkevitch and J. C. Autran Prediction of Baking Quality of Bread Wheats in Breeding Programs by Size-Exclusion High-Performance Liquid Chromatography. Cereal Chem. (1989) 66(6):448:456, 1989.

[8] J. Garrin Fullington, E. W. Cole, D. D. Kasarda. Quantitative SDS–PAGE of total protein from different wheat varieties. J. Science of food and Agric. (1980) Volume 31, Issue 1, pages 43–53.

[9] P. C. Williams, K. H. Norris and D. C. Sobering. Determination of protein and moisture in wheat and barley by near infrared transmission. *J. Agric. Food Chem.* (1985) *33* (2), pp 239–244.

[10] S. R. Delwiche, R. A. Graybosch and J. Peterson. Predicting Protein Composition, Biochemical Properties, and Dough-Handling Properties of Hard Red Winter Wheat Flour by Near-Infrared Reflectance. Cereal Chem.(1998) 75(4):412-416.

[11] J. Blažek, O. Jirsa And M. Hrušková Prediction of Wheat Milling Characteristics by Near-Infrared Reflectance Spectroscopy. Czech J. Food Sci. Vol. 23 (2005), No. 4: 145–151.

[12] E. B. Maghirang and F. E. Dowell. Hardness Measurement of Bulk Wheat by Single-Kernel Visible and Near-Infrared Reflectance Spectroscopy. Cereal Chem. (2003) 80(3):316-322.

[13] S. R. Delwiche. Measurement of single-kernel wheat hardness using near-infrared transmittance. Transactions of the ASABE. (1993) 36(5): 1431-1437.

[14] B. G. Osborne and T. Fearm. Collaborative evaluation of near infrared reflectance analysis for the determination of protein, moisture and hardness in wheat. Journal of the Science of Food and Agriculture Volume 34 (1983), Issue 9, pages 1011–1017.

[15] H. L. Mark and D. Tunnell. Quantitative near-infrared reflectance analysis using Mahalanobis distance. Anal. Chem. (1985) 57:1449-1456.

[16] K. C. Lawrence, W. R. Windham and S. O. Nelson. Wheat Moisture Determination by 1-to 110-MHz Swept-Frequency Admittance Measurements. Transactions of the ASAE (1998) vol. 41(1):135-142.

[17] S. R. Delwiche and R. A. Graybosch. Identification of Waxy Wheat by Near-infrared Reflectance Spectroscopy. J. of Cereal Science **35** (2002) 29–38.

[18] S. R. Delwiche. Single Kernel protein content in wheat by near-infrared reflectance. ASAE (1996) 96-3031.

[19] S. Mahesh, D. S. Jayas, J. Paliwal and N.D.G. White. Identification of western Canadian wheat classes at different moisture levels using near-infrared (NIR) hyperspectral imaging. The Canadian Society for Bioengineering (2008) paper No. 08-196.

[20] C. B. Singh, D. S Jayasa, J. Paliwala and N.D.G. White. Identification of insect-damaged wheat kernels using short-wave near-infrared hyperspectral and digital colour imaging. Computers and Electronics in Agriculture, Volume 73 (2010), Issue 2, Pages 118–125.

Conclusions

1. Chapter 2. Development of robust soybean NIR calibration models with high variability and temperature compensation in the base data

. The use of hierarchical models with multivariate methods for the reduction of large data sets, dismissed the noise and improved the ability of accuracy and prediction of the predicted values.

. The inclusion of temperature compensation samples into the base model showed slight improvements in the models, but reasonable for future reliable predictions.

2. Application of Near Infrared Spectroscopy technology and hyperspectral NIR imaging for the detection of fungicide treatment on durum wheat samples

. The results obtained in this study showed that the application of Near Infrared Reflectance Spectroscopy and Hyperspectral NIR imaging could be useful and fast method to use for the assessment of durum wheat disease effects.

3. Application of NIRS in authentication of bread wheat varieties from southern Spain.

. The potential of Near Infrared for the determination of bread wheat samples according to their varieties is showed in this study. The results obtained showed the capacity of NIR to varietal discrimination.

Annexes

Talanta 97 (2012) 298–302

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Application of near infrared spectroscopy technology for the detection of fungicide treatment on durum wheat samples

M. Soto-Cámara *, A.J. Gaitán-Jurado, J. Domínguez

IFAPA Alameda del Obispo, Production Area, Avda Menéndez Pidal s/n, 14004 Córdoba, Spain

ARTICLE INFO

Article history:
Received 27 September 2011
Received in revised form
13 April 2012
Accepted 19 April 2012
Available online 12 May 2012

Keywords:
NIR
Durum wheat
Puccinia triticina
Septoria tritici
Fungicide
Discrimination

ABSTRACT

The feasibility of Near Infrared Spectroscopy to detect fungicide treatment on wheat samples was assessed. A total of 213 durum wheat samples from four different trial sites in Andalusia (southern Spain), with different agroclimatic conditions (soil, temperature, rainfall) were selected for been analyzed on VIS+NIR (400 nm–2500 nm) and NIR (1100 nm–2500 nm). Different mathematical pre-treatment on the signal (scatter correction and derivatives) were evaluated for their discrimination accuracies. Using MPLS, the selected models obtained 84% of well classified samples.

1. Introduction

Durum wheat is an important crop in the Mediterranean area. The main uses are in human food products, like bread, pasta and couscous [1].

Andalusia is the leading region producing durum wheat in Spain. It contributes more than 74% of the total national production [2].

Wheat genotypes, agronomics conditions and fertility inputs are the foremost factors determining durum wheat yield and quality characteristics [3]. Nevertheless, an important bounding aspect of durum wheat is the damage caused by diseases. Two of the most important are, above all, leaf rust and septoria leaf spot (incited by *Puccinia triticina* and *Septoria tritici*, respectively). Plant diseases are greatly influenced by environmental factors, including known stresses as deficiencies of essential nutrients and/or toxicities of other mineral elements [4]. Modifications in cultural practices, such as direct sowing, use of Nitrogen fertilizers and irrigation, may contribute to an increase on the disease severity [5]. *Puccinia triticina* is the most common rust of wheat. It has affected wheat for thousands of years. Yield losses in wheat from *Puccinia triticina* infections are usually the result of decreased numbers of kernels per head and lower kernel weights [6].

Methods used to fight fungal diseases and in the development of new fungicides in cereals, are based on etiological and epidemiological knowledge. The presence of a particular fungal disease is related to the degree of susceptibility of the variety, presence of inoculum, plant phenological status and climatological factors, especially those associated with humidity [7].

When infective fungus parts get accumulated on the grain surface; enzymes destroy proteins, starch granules and grain cell walls [8].

Nowadays, consumers are more conscious of eating high quality products free of toxic agents. Increased food scrutiny requires the development of improved and more readily available analytical methods for food products authentication and detection of contaminant [9–12]. Near Infrared Spectroscopy (NIRS) has been used for the determination and quantification of proximate quality parameters on food (protein, fat, sugar) and for the recognition of transgenic foods [13].

The objective of this work was to evaluate NIR technology to detect differences between durum wheat seed samples coming from plants which have been treated with fungicide and those coming from non treated plants, using discrimination models.

2. Materials and methods

2.1. Experimental design

All the durum wheat samples came from trials carried out on randomised complete block designs with four replications. It is the most common design used in field trials. Crop management

* Corresponding author. Tel.: +34 957016070.
E-mail address: mariana.soto.ext@juntadeandalucia.es (M. Soto-Cámara).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.04.034

on trials was the standard used by farmers on the area. Experimental plots were 12 m^2 (10 m × 1.20 m).

2.2. Wheat samples

The 213 wheat samples from the 27 durum wheat varieties were provided by the Andalusia Network of Agrarian Experimentation (RAEA), managed by the Andalusian Institute of Agricultural Research and Training (IFAPA). Samples originated from four different trial sites located in Jerez (Cadiz), Camino de Purchil (Granada), Tomejil (Seville) and Santaella (Cordoba), each having quite different agroclimatic conditions and grown 2009–2010 seasons, were used.

Durum wheat seeds were sent to IFAPA Alameda del Obispo, Cordoba, in paper bags containing approximately 500 g, and then they were kept on small 250 g plastic containers. Two samples of every variety were received, one, coming from fungicide treated plants (T) and the other from plants free of it (O), 105 and 108 sample of each respectively.

2.3. Fungicide treatment

Fungicide treatment against leaf diseases, as the leaf rust Roya and Septoria leaf spot (incited by *Puccinia triticina* and *Septoria tritici*, respectively) consisted of one application with a concentrated suspension of 12.5% p/v of epoxiconazol, Lovit (Basf España S.A.).

A dose of 1 L ha^{-1} (the maximum recommended dose) was applied at the phenological phase of flag leaf unfolded. When the crop was in stage 39 of the code of Weber and Bleiholder [14], flag leaf stage: flag leaf fully unrolled, ligule just visible [15–16]. The security time limit of 42 days before harvest was followed.

2.4. Chemical analysis

Traditional reference methods were used to compare the quality parameters of both groups of samples. For total content of crude protein the Kjeldahl method was used (Panreac B.O.E 19-7-1977 and 20-7-1977). Moisture was determined by the Panreac air oven method (B.O.E 19-7-1977 and 20-7-1977). Gluten index in order to determine water insoluble protein the Panreac official method was followed (B.O.E 19-7-1977 and 20-7-1977). Finally, total weight of 1000 wheat kernels was performed with Numigral I.

2.5. NIR spectra

The spectra were recorded on a *Foss NIRSystems* (model 6500 Foss-NIRSystems, Inc., Silver Spring, MD, USA) in reflectance mode, over a wavelength range between 400 and 2500 nm (Visible and NIR region), measured in a 2 nm steps.

Intact grain samples were placed in a cuvette of 16.5 × 3.5 cm, with a quartz window (Sample Cell NR-7080) showed in Fig. 1. Two spectra per sample were obtained. The cell was filled with about 20 g, scanned and finally returned to the container and mixed with the remaining sample (approximately 480 g). The process was repeated again to obtain the second sample spectrum. To avoid packing variations, only one analyst did sample preparation.

Optical density was stored as log (1/R), where R is the reflectance energy recovered by a split detector system with silicon (Si) between 400 nm and 1098 nm and a lead sulfide (PbS) between 1100 nm and 2500 nm. A personal computer with the software ISIscan (v2.81; Infrasoft International LLC. Port Matilda now State College, PA, USA) was used for the operation of the spectrometer, and to store and manage optical data.

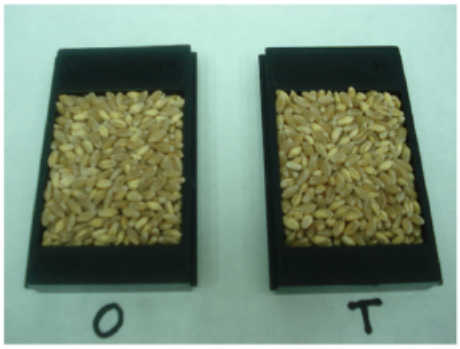


Fig. 1. Display of wheat intact grains in the cuvette. Wheat with and without treatment (T and O).

2.6. Statistical analysis and discriminant equations

2.6.1. Root Mean Squared (RMS)

Filtering of the subsamples spectra was done by calculating the RMS for this sample presentation form.

The following expressions were applied to calculate RMS values:

$$RMSj = \sqrt{\frac{\sum_{i=1}^{n}(Y_{ij}-\overline{Y}_i^2)}{n}}$$

$$STD = \sqrt{\frac{\sum_{j=1}^{n}(RMSj)^2}{n-1}}$$

$$STD_{Limit} = 1.036x\sqrt{\frac{\sum_{k=1}^{k=m} STD_k^2}{m}} = 1.036x\sqrt{\overline{STD}^2}$$

where n is the number of data (absorbance readings), m is the number of samples, Y_{ij} is the absorbance value log (1/R) for subspectrum j at wavelength i (λ_i) and $\overline{Y}$ is the absorbance value log (1/R) for the average spectrum of a sample at wavelength i (λ_i).

The STD_{Limit} (Standard Deviation Limit) values were used to obtain RMS_{Limit}. Once the spectra of samples that exceeded the cut-off limit were eliminated, other spectra were obtained on the same sample, obtaining new RMS values. If the new RMS value exceeded the limit again, this sample was marked as not suitable to be included in the calibration set [17].

2.6.2. Calibration and validation sets

The sample set used in the study (185 samples after RMS) was split into a calibration set containing 158 samples (80% of the total) and a validation sample set comprising 25 samples (20%). This splitting was carried out in a random way by WinISI software.

WinISI III (v1.50e, Infrasoft International LLC) software was used for spectral data analysis and development of chemometric models.

2.6.3. Principal Component Analysis (PCA)

Prior to classification models, Principal Components Analysis (PCA), an orthogonal transformation that enables a subspace of R^d to be obtained with a minimum loss of information [18] was carried out. Twenty-four Math Pretreatments were applied to spectra to develop the PCA as a result of combinations of derivative (0, 0, 1; 1, 4, 4; 2, 4, 4; 2, 10, 5), scatter correction (SNVD, Standard Normal Variate and Detrend; MSC, Multiplicative Scatter Correction) and spectral range (VIS–NIR; NIR).

The qualitative difference between varieties and removal outliers was done with a standardized Mahalanobis (GH) distance. Distances between each sample and the population center greater than 3 are marked as possible spectral outlier [19]. A sample was eliminated if it appeared as anomalous repeatedly in the different math treatment mentioned.

2.6.4. Partial least squares modified (MPLS)

The discriminant model was built using Modified Partial Least Squared (MPLS), in winISI software; this assigns to each spectrum a value called a "dummy" variable (or discriminant variable). The new variable obtained, acquired a value between 1 (samples in the group with no fungicide treatment) or 2 (with fungicide treatment group). The discriminant variable limit established for group selection was ≥ 1.5 [20]. Which means: samples with a value < 1.5 is included in one group (in this case are O samples) and samples with a high value of 1.5 belong to the other group.

A maximum of twelve PLS terms were selected, if the model selected the 12 PLS terms, the process was repeated with two more terms to avoid overfitting effect. Internal cross validation (with five cross validation groups) was used in order to estimate the final number of PLS terms. Using a cross validation with five groups, on the first pass, the samples of group 1 are used for the validation, and those of the remaining four groups are used for the actual calibration. In pass 2, group 2 is used for the actual calibration; in pass 3, group 3, and so on [21]. The Math treatments used in both cases were the same as applied in PCA analysis.

The criteria used to select the best models were: Coefficient of determination of calibration (R^2), Standard Error of Cross-Validation (SECV) and % of samples correctly classified.

3. Results and discussion

3.1. Prior analysis

When Fungus leave infection is produced early in the season, often prevents the development of the grain. When foliar fungus infection appear, the photosynthesis area decreases which normally means a drop in the amount of protein, starch and kernels size and weight. However, once the grain has been filled, and subsequently infection occurs, its size will not be affected, even the color and appearance.

Fig. 2 represent the mean spectra [log 1/*R*] of *T* and *O* samples. In the 1200–1318 nm, 1440–1880 nm and 1920–2498 nm wavelength regions there were small differences which might be due to different chemical composition inn wheat samples. Burns [22] reported that the 1940 nm peak is related with moisture in flour, while 2180 nm was assigned to protein absorption, 2150 nm area has been used for protein [23] and for starch 2100 nm. Delwiche and Hareland [24] indicated that the region 1130–1190 nm was a stable region for defining a difference that could be used in classifying normal and scab damaged kernels.

3.2. Reference analysis

At first sight there is no evidence of any modification in kernels size or color by fungus infection. In order to find kernels matrix differences, chemical analysis were performed. There were not important differences between both groups in moisture or gluten index, being both groups compensated in a similar percentage. On the other hand, in protein and 1000 wheat kernels weight showed remarkable differences.

Figs. 3 and 4 represent the difference between T and O samples. When the difference appears in the positive part means that the

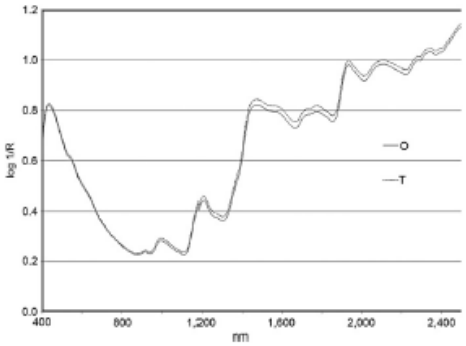


Fig. 2. Average spectra of treated (-) and untreated (- -) samples.

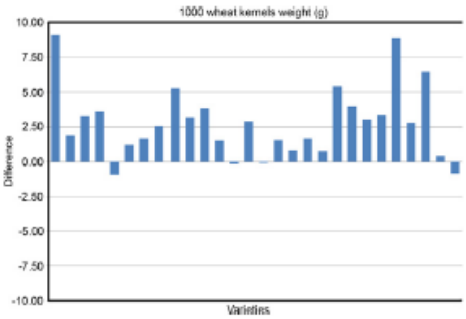


Fig. 3. Difference between T and O samples in weight of 1000 wheat kernels (grams).

measure of T samples is higher than in O samples. As can be seen there is an evident dominance in both parameters in T samples.

Thus, the parameters protein and thousand grain weight (indicators of the quality of the grain) were higher in treated samples (*T*), mainly due to the effect that caused the fungicide that prevents the onset of disease.

3.3. RMS

In order to obtain the RMS cut-off value, the method described in Section 2.5 was followed. An individual value for each sample was calculated setting a maximum value (RMS cut off) of 10000. There was a tiny difference between the limit value of T and O samples with a value of RMS limit of 8000 and 9000 respectively.

Samples which surpassed the RMS cut-off limit were scanned again following the steps detailed in Section 2.5. Finally a total of 12.5% (14 samples) of *T* and 14.7% (16 samples) of *O* were eliminated from whole group.

3.4. PCA

To extract initially spectra information and qualitative differences between all samples, data analysis was carried out applying the limit criteria discussed in Section 2.4. After all the mathematical treatments, only one spectrum was eliminated (T sample). WinISI program picked 9 factors to cover 99.97% of the explained variance.

Before MPLS analysis of the calibration group the accumulative reliabilities of the first 3 PCs were 99.18%, the fourth PC contributed an additional 0.45% of the total variance, the fifth 0.10% and 0.24% the remaining four. This means that the most important information and spectra features were included in the three first PCs.

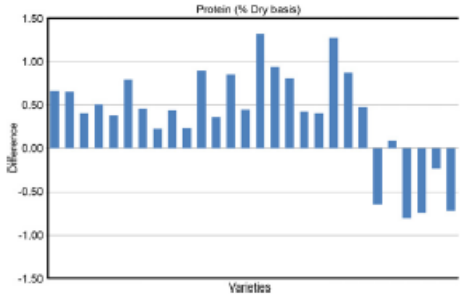


Fig. 4. Difference between T and O samples in wheat % protein (dry basis).

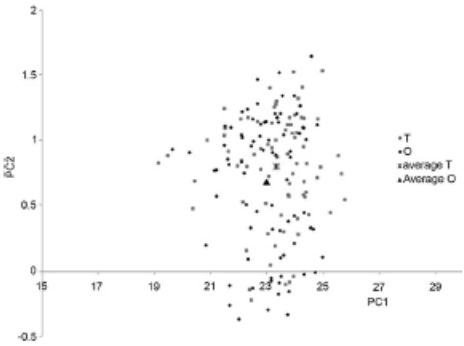


Fig. 5. Principal Component Analysis: first PC versus second PC on T (treated) and O (non treated) samples.

Using the PCA scores, there was no separation between T and O groups and their corresponding centroids (Fig. 5).

3.5. MPLS

MPLS discriminant models were developed. The dummy variable was set as a referent values for the O and T group. The specification set was 1 for O samples and 2 for T samples.

A maximum of 12 PLS factors and 5 groups of cross validation were used in all the PLS models. After all cross validation passes and with the individual statistical parameter of each group, the number of factors for the smallest error is established. The number of factors required on the spectral MPLS analysis was 7–8 depending on the mathematical treatment applied.

A blind test with 25 samples (12 T and 13 O) was carried out in order to obtain an external validation. With appearance not difference between both groups of samples (T and O samples), high percent of correctly classified were obtained. Table 1 shows the best results obtained on the reflectance mode in NIR and VIS+NIR regions. The best discriminant model was that obtained using the derivative treatment MSC 2,4,4, which displayed a SECV of 0.34, R^2 of 0.78 and with an external validation of 84% of samples correctly classified. Despite those final values of the models for both regions result very similar, higher R^2 and lower SECV were obtained when the VIS was included.

Lou et al. [25] and Chandra et al. [26] developed discrimination models to assign categories of wheat kernel damage using a color machine and hyperspectral image. Their results were similar to those of the Foss NIR system 6500.

Fig. 6 shows a linear representation of the external validation. Quadrant T (T samples) displays one sample belonging to group O; this sample could present a higher resistance to fungus infection which means less modification of the chemical composition. On the other hand, quadrant O presents three samples misclassified, samples coming from treated plants but appearing in the region on untreated samples. Menniti et al. [27] reported the ineffectiveness of epoxiconazol at controlling some other fungal diseases, which could also be the case in the plants from which these samples come from.

T samples, including as in the group of O, can be due to a failure of treatment which is not 100% effective and therefore has caused disease in some plants. Conversely, the O samples included in the T group may be because some varieties or growing conditions have allowed disease does not develop in some untreated plants, which are included in the treated group. In addition to these justifications must be added the error of the model itself.

Table 1
Parameter values of the best models developed for VIS+NIR and NIR regions.

Spectral range	*Math treatment*			*Classification matrix*				*External validation*
	Scatter	*Derivative*	*PLS factors*	O	T	R^2	*SECV*	(% Correctly classified)
VIS+NIR (400–2500 nm)	None	2, 4, 4	8	**61** 19	16 **62**	0.76	0.35	0.84
	SNV+DT	2, 4, 4	7	**65** 15	9 **69**	0.74	0.35	0.84
	MSC	2, 4, 4	8	**67** 13	10 **68**	0.78	0.34	0.84
NIR (1100–2500 nm)	None	2, 4, 4	8	**67** 13	14 **64**	0.71	0.37	0.80
	SNV+DT	2, 4, 4	12	**67** 13	16 **62**	0.84	0.39	0.76
	MSC	2, 4, 4	8	**66** 14	14 **64**	0.72	0.36	0.80

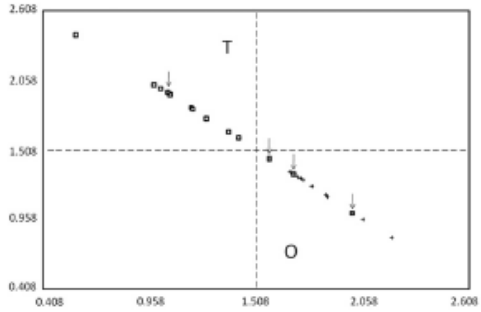


Fig. 6. Samples misclassified on the external calibration group.

4. Conclusions

The results obtained in this study showed that the application of Near Infrared Reflectance Spectroscopy could be a useful and fast method to use for the assessment of fungicide treatment on durum wheat. We developed a model using 158 samples of durum wheat seeds coming from plants with and without fungicide treatment. This model displayed an accuracy of 84%. These results were obtained using samples from only one crop season. Currently we are in the process of obtaining models applied to further cropping seasons to assess the reproducibility of the method.

References

[1] ⟨www.fao.org⟩, 2010.
[2] Anuario de Estadística Ministerio de Medio Ambiente y Medio Rural y Marino. Marm, 2009.
[3] F.M. Dupont, S.B. Altenbach, J. Cereal Sci. 38 (2003) 133–146.
[4] Nand Kumar Fageria, Virupax C. Baligar, Charles Allan Jones, Growth and Mineral Nutrition of Field Crops, Third Edition, CRC Press, 2011.
[5] Z. Eyal, A.L. Scharen, J.M. Prescott, M. van Ginkel, Enfermedades del trigo causadas por Septoria: Conceptos y métodos relacionados con el manejo de estas enfermedades, CIMMYT, México, D.F., 1987 45p.
[6] Melvin D. Bolton, James A. Kolmer, David F. Garvin, Mol. Plant Pathol. 9 (5) (2008) 563–575. 2008.
[7] López Bellido L., Cultivos herbáceos, vol. I, Cereales. Ed Mundi-Prensa, Madrid, Spain, (1991) 539 pp.
[8] ⟨www.alimentacion.org.ar⟩. 2010.
[9] R.M. Balabin, E.I. Lomakina, Analyst 136 (1703) (2011) 2011.
[10] R.M. Balabin, R.Z. Safieva, Fuel 87 (2745) (2008) 2008.
[11] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Microchem. J. 98 (121) (2011) 2011.
[12] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Anal. Chim. Acta 671 (27) (2010) 2010.
[13] A. Alishahi, H. Farahmand, N. Prieto, D. Cozzolino Identification of transgenic foods using NIR spectroscopy: A review. Spectrochim. Acta Part A 75 (2010) 1–7.
[14] E. Weber, H. Bleiholder, Gesunde Pflanzen 42 (1990) 308–321.
[15] Uwe Meier. Estadios de las plantas mono- y dicotiledóneas. BBCH Monografía, Centro Federal de Investigaciones Biológicas para Agricultura y Silvicultura, 2001.
[16] Mar Cátedra Cerón, I. Solís Martel, Span. J. Agric. Res. 1 (3) (2003) 19–26.
[17] Antonio J. Gaitán-Jurado, María García-Molina, Francisco Peña-Rodríguez, Víctor Ortiz-Somovilla, J. Near Infrared Spectros. 16 (2008) 421–429.
[18] Y. Langeron, M. Doussot, D.J. Hewson, J. Duchêne, Classifying NIR spectra of textile products with kernel methods, Eng. Appl. Artif. Intell. 20 (3) (2006).
[19] J.S. Shenk, M.O. Westerhaus, Crop Sci. 31 (1991) 1548–1555.
[20] G. Downey, Discrimination et authentification des aliments et des ingredients alimentaires par spectroscopie dans l'infrarouge proche et moyen, in: D. Bertrand, E. Dufour (Eds.), La spectroscopie Infrarouge Et Ses Applications Analytiques, Editions TEC & DOC, Paris, France, 2000, pp. 397–422.
[21] ⟨www.zeiss.com⟩. 2011.
[22] A. Burns, Emil W. Ciurczak, Handbook of Near-Infrared Analysis, third edition, CRC Press, 2008 2008.
[23] Ana Ruano-Ramos, Antonia García-Ciudad, Balbino García-Criado, J. Sci. Food Agric. 79 (1999) 137–143.
[24] Stephen R. Delwiche, Gary A. Hareland, Cereal Chem. 81 (5) (2004) 643–649.
[25] X. Lou, D.S. Jayas, S.J. Symons, J. Cereal Sci. 30 (1999) (1999) 49–59.
[26] Chandra B. Singh, Digvir S. Jayas, Jitendra Paliwal, Noel D.G. White, Biosyst. Eng. 105 (2010) 380–387.
[27] Anna Maria Mennitti, Davide Pancaldi, Massimo Maccaferri, Lucia Casalini, Eur. J. Plant Pathol. 109 (2003) 109–115.

Detección de variedades de trigo harinero (*Triticum aestivum* L.) mediante Espectroscopía de Infrarrojo Cercano

Soto M., Gallán-Jurado A. J., Font R., Del Río-Celestino M., Salguero L., España F. y Domínguez J.
IFAPA- Centro Alameda del Obispo. Apartado 3092, E-14080 Córdoba, España.

Resumen

El incremento por la necesidad de un producto de calidad y de la seguridad alimentaria se ha hecho visible a lo largo de los años;, esto ha convertido dichos factores en objetivo principal para la Industria Agroalimentaria. Una de las herramientas analíticas más útil, rápida y limpia empleadas para el Control de la Calidad en la Industria es la Espectroscopía de Infrarrojo Cercano; durante años se ha evaluado su capacidad para predecir parámetros de calidad importantes (proteínas, fibra, humedad...).obteniéndose resultados fiables y precisos.

En la industria panadera existen variedades de trigo de distinta calidad, esto influye tanto en el resultado de sus productos como en su valor comercial, con lo cual el objetivo del presente trabajo es el estudio de la viabilidad de la tecnología NIR como método para discriminar variedades de trigo.

Materiales y métodos

Un total de 154 muestras de trigo harinero (*Triticum aestivum* L.) , proporcionadas por la RAEA (Red Andaluza de Experimentación Agraria) y pertenecientes a cinco variedades ("Cartaya", "Odiel", "Yecora", "Galeón" y "Gazul") diferentes por su calidad panadera, se emplearon en el trabajo. Todas las variedades fueron cultivadas en diferentes localidades de Andalucía, con diferentes condiciones climáticas y edáficas a lo largo de cuatro años (2004-2008). Fueron escaneadas con un espectrofotómetro NIR (modelo 6500 Foss- NIRSystems, Inc., Silver Spring, MD, USA) en el módulo de reflectancia desde la región de los 400 hasta los 2500 nm y recogida como el log(1/R) a 2 nm de intervalo. El software que se usó para el desarrollo de los modelos matemáticos (PLS) fue WinISI III versión 1.5.

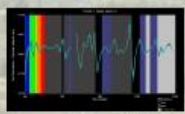
Como etapa previa al desarrollo de los modelos cualitativos, se procedió a realizar una análisis de componentes principales (ACP). Los tratamientos matemáticos seleccionados fueron: Scatter: none, SNV+D y WMSC derivación 0,0,1, 1,4,4 y 2,15,8 (Miralbes, 2008) tanto en la región VIS-NIRS como en la NIRS solamente.

El análisis discriminante desarrollado se basó en una regresión por mínimos cuadrados parciales modificados(MPLS), se empleó un grupo de calibración de 113 muestras y uno de validación de un total de 28 muestras.

Foss NIRSystem 6500, módulo transport.

Scatter:None, SNV+D y WMSC

Derivación: 0,0,1, 1,4,4 y 2,15,8



Representación espectral del conjunto de muestras en segunda derivada.

Resultados y discusión

Al aplicar todos los tratamientos matemáticos se observó que los mejores resultados obtenidos eran con Scatter: WMSC y una derivación 2,15,8 lo que dieron lugar al modelo discriminante de mayor potencial predictivo con un 95,5% de certeza, cuyos valores aparecen en la tabla 1.

El mayor error de clasificación obtenido fue del 4,5%, un total de 5 muestras mal clasificadas de las 111 usadas para el grupo de calibración. Con un SECV bajo 0,2262 y un R^2 de 0,6624 que nos permite una buena separación entre los valores bajos, medios y altos (Garrido, A. Et al 2009)

En el grupo de validación se usaron un total de 36 muestras dando como resultado un error de clasificación del 5,55%.

Scatter	Trat. deriv.	PLS	Error Clasif.	SECV	R^2
None	2,15,8	9	4,5%	0,23	0,64
SNV+D	2,15,8	9	4,5%	0,23	0,64
WMSC	2,15,8	9	4,5%	0,23	0,66
None	1,4,4	10	18,01%	0,28	0,49
SNV+D	2,15,8	10	13,51%	0,28	0,51
WMSC	1,4,4	10	13,51%	0,28	0,50

VIS-NIRS
NIRS

	Cartaya	Galeón	Gazul	Odiel	Yecora
Cartaya	25	0	1	0	0
Galeón	0	24	0	1	0
Gazul	0	0	30	0	0
Odiel	0	0	0	10	1
Yecora	1	0	1	0	17

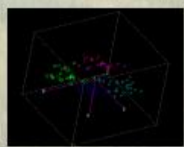


Tabla 1. Resultado del análisis discriminante.

Conclusiones

Los resultados obtenidos en el presente trabajo demuestran que pequeñas variaciones (físicas o químicas) que hacen diferir entre las variedades de trigo son detectados por esta tecnología, permitiendo una buena discriminación y una correcta clasificación entre variedades.

Instituto de Investigación y Formación Agraria y Pesquera
CONSEJERÍA DE AGRICULTURA Y PESCA

www.grainlab.org
Grain Quality LABORATORY
IOWA STATE UNIVERSITY

15th International Conference on Near Infrared Spectroscopy
NIR 2011
13 - 20 MAY 2011

Instituto de Investigación y Formación Agraria y Pesquera
CONSEJERÍA DE AGRICULTURA Y PESCA

Development of Robust Calibrations From Large Databases

Mariana Soto-Cámara[1], Lidia Esteve Agelet[2], Nanning Cao[2], Charles R. Hurburgh[2], Glen Rippke[2], Juan Dominguez-Giménez[1], and Antonio J. Gaitán-Jurado[1]
[1]IFAPA Alameda del Obispo, Avda. Menéndez Pidal, s/n, PO Box 3092, 14080 Córdoba, Spain
[2]Department of Agriculture and Biosystems Engineering, Iowa State University. Ames, Iowa 50010, USA
*Corresponding author: mariana.soto@juntadeandalucia.es

Introduction:

When using a large calibration database, selecting and retaining the most representative samples for calibration development becomes essential. Retaining too many samples may not bring any additional benefit but can introduce redundancy and incorporate unnecessary noise. The objective of this study was to build robust models for **soybean protein**, selecting the optimal data set from a large pool of samples from different crop years.

Materials and methods

Five Bruins NIR instruments, Puchheim, Germany, (OmegaAnalyzer G 106110 and G 106118, G 609448, Omega S 21101 and Agri Check 31002) and 9 crop seasons (2001 – 2009) were used for calibration.

The data was organized in Unscrambler 9.8 (Camo, Oslo, Norway). and the general models were made in Unscrambler. MatLab 2010A, (MatWorks, Inc., Natick, MA, USA.) PLS toolbox, (Eigenvector Research, Inc. Wentachee, WA., USA) was used for the uniform distribution data models.

The validation set was a sample group from the 2010 crop.

Outliers (spectral and chemical) were removed. The models were built two ways:

1. Development of a calibration from the complete data base (all instruments and available years.

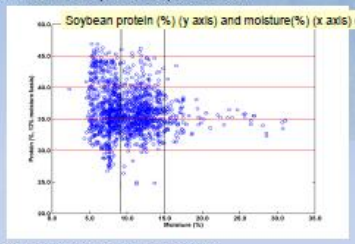
Instr/year	2001	2002	2003	2004	2005	2006	2007	2008	2009
6110	112	131	112	5	143	148	167	148	121
6118	169	143	170	163	135	5	149	147	172
21101	0	131	178	148	141	157	164	142	120
31002	0	0	0	0	0	0	165	18	115
609448	0	0	0	0	0	0	110	100	65

2. The database was split into quadrants. Protein was the main factor and moisture was considered for reality.

Moisture: Low moisture (below or equal 8.9%), medium (9.0% to equal 14.9%), and high (>15%).

Protein increments of 5% (<29.9%, equal 30.0% to up to 34.9%, equal 35.0% and up to 39.9%, equal 40.0% to 44.9%, and equal or above 45%)

Data was taken randomly from those quadrants ensuring uniform distribution, using Matlab, of both protein and moisture. Quadrants that had small amount of samples, all samples were taken.



100 times of each of these combinations:
1. Randomly select 200 max per each quadrant → 1528 samples
2. Randomly select 100 max per each quadrant → 938 samples
3. Randomly select 50 max per each quadrant → 520 samples
4. Randomly select 25 max per each quadrant → 289 samples
5. Randomly select 10 max per each quadrant → 126 samples

3. The validation set used for all models are from the 2010 samples:

Instrument	samples
6110	76
6118	77
21101	77
31002	72

Bruins OmegAnalyzerG

BRUINS INSTRUMENTS

Bruins Instruments
Lindberghstrabe 12
82178 Puchheim, Germany
Phone +49-89-800-677-0
Email: support@bruins.de

Results

Calibration results with all instruments, all years and all samples, soybean protein, Bruins Omega.

	N	Factor	R^2	RMSE(%)	SEP(%)	Bias
Calibration	4093	10	95.9%	0.71		
Validation	302	10	96.2%	0.81	0.76	0.23

Validation results of 100 models (iterations) per each sample size (N) with 10 Factors (constant).

Combination	N Calibration	Average R^2	SEP(%)				Bias(%)			
			Min	Max	Average	Std	Min	Max	Average	Std
1	1528	96.6%	0.76	0.83	0.80	0.02	-0.20	-0.31	-0.25	0.02
2	938	93.4%	0.78	0.89	0.82	0.02	-0.18	-0.35	-0.27	0.03
3	520	96.3%	0.76	0.96	0.84	0.03	-0.18	-0.40	-0.27	0.03
4	289	96.1%	0.78	1.00	0.86	0.04	-0.10	-0.43	-0.28	0.07
5	126	95.3%	0.82	1.26	0.96	0.08	0.03	-0.64	-0.30	0.14

Bruins AgriCheck

Conclusions

From a large data set, an aggressive selection of samples for calibration allowed building robust models. Using about 500 samples gave results similar to models for the same instrument using over 3000 samples.

The SEP increases with smaller data sets; 300-600 samples appeared optimum.

Written by:

Mariana Soto Cámara