University of Modena and Reggio Emilia
Ph.D. School in AgriFood Sciences, Technologies
and Biotechnologies

"Coupling FT-NIR spectroscopy and multivariate analysis for fast and non destructive characterization of bread wheat and swine adipose tissue"

Ph.D. student: Davide Salvo

XXVI cycle

Tutor: Dott.ssa Giorgia Foca
Co-Tutor: Prof. Alessandro Ulrici

Dean of Ph.D. School: Prof. Andrea Pulvirenti

Titolo: "Accoppiamento di spettroscopia FT-NIR ed analisi multivariata per la caratterizzazione rapida e non distruttiva di grano tenero e tessuto adiposo suino"

Riassunto

La spettroscopia nel vicino infrarosso (NIR) è una tecnica analitica rapida, non distruttiva, economica e precisa che ha trovato applicazioni nel campo dell'industria alimentare, grazie alla sua capacità di fornire informazioni qualitative e quantitative anche senza preparazione del campione. Sebbene le attuali applicazioni della spettroscopia NIR nella caratterizzazione di matrici alimentari complesse siano numerose, ulteriori sviluppi, spesso associati all'uso di metodi chemiometrici innovativi, potrebbero essere di grande valore per l'industria alimentare.

La presente ricerca si è focalizzata sulla caratterizzazione di due matrici alimentari, frumento e tessuto adiposo suino, mediante spettroscopia NIR accoppiata a metodi chemiometrici.

Nel caso del frumento e derivati (farine bianche, sfarinati e granelle) è stata studiata la possibilità di utilizzare la spettroscopia NIR per prevedere la qualità delle materie prime da panificazione. La classificazione delle farine in categorie qualitative viene effettuata da valutatori esperti sulla base dell'Indice Sintetico di Qualità (ISQ), che viene assegnato ai campioni di frumento in base alle loro proprietà reologiche e chimiche. Tuttavia, la procedura non è semplice e spesso l' assegnazione alle classi risulta incerta, facendo sorgere la possibilità di controversie. Inoltre, la determinazione dei parametri ISQ richiede analisi di laboratorio lunghe e costose. Per risolvere questo problema, è stato applicato l'algoritmo di calibrazione Partial Least Squares (PLS) agli spettri NIR acquisiti sui campioni in tre diverse forme fisiche, al fine di prevedere quantitativamente i valori delle proprietà ISQ. Utilizzando diversi tipi di approcci PLS (PLS1 e PLS2) si è dimostrata la correlazione tra gli spettri NIR ed i parametri ISQ, ottenendo risultati promettenti nella determinazione del peso ettolitrico e del parametro reologico W. Per quel che riguarda la caratterizzazione del tessuto adiposo suino, la ricerca si è focalizzata sullo sviluppo di metodi rapidi ed economici per classificare due strati sottocutanei aventi differente composizione in acidi grassi e destinati alla produzione di alimenti diversi. A tal scopo, i campioni di grasso sono stati analizzati mediante colorimetria, spettroscopia NIR, e analisi distruttive tradizionali quali la determinazione del numero di iodio (IV) e misurazioni gas cromatografiche mirate alla determinazione della percentuale di acidi grassi specifici come pure i loro importi globali in termini di acidi grassi saturi (SFA) e acidi grassi polinsaturi (PUFA). La classificazione dei due strati è stata ottenuta mediante l'uso dell'Analisi Discriminante PLS (PLS-DA) applicata ai dati colorimetrici e agli spettri NIR. L'analisi degli spettri ha anche previsto l'utilizzo di vari pretrattamenti e di algoritmi di selezione di variabili. I risultati in classificazione hanno confermato la possibilità di distinguere i due strati di tessuto adiposo: mentre sui dati colorimetrici si è ottenuta un'efficienza in classificazione del 78% , sui dati NIR si è raggiunta un'efficienza pari al 96%. Per quanto riguarda le analisi quantitative, i modelli di calibrazione ottenuti hanno mostrato elevate capacità predittive per la stima del grado di insaturazione rappresentato dall'IV, per acido linoleico, SFA e PUFA, con coefficienti di correlazione in predizione attorno a 0.8; gli algoritmi di selezione di variabili hanno prodotto lievi miglioramenti nei risultati.

Parole chiave:
Spettroscopia FT-NIR
Chemiometria
Alimenti
Frumento
Tessuto adiposo suino

Title: "Coupling FT-NIR spectroscopy and multivariate analysis for fast and non destructive characterization of bread wheat and swine adipose tissue"

Abstract

Near InfraRed (NIR) spectroscopy represents a rapid, non-destructive, cheap, and accurate technique that has found increasing applications in the field of food industry thanks to its ability of providing both qualitative and quantitative information with minimal or no sample preparation. Despite the current applications of NIR spectroscopy in the characterization of different food matrices, further developments concerning new applications as well as the use of chemometric methods for the extraction of useful information could be highly valuable for the food industry.

In this context, the proposed research has been focused on the characterization by means of NIR spectroscopy coupled with chemometrics of two main food matrices: wheat and swine adipose tissue.

As for wheat, the possibility to use NIR spectroscopy for the prediction of quality of raw matters has been investigated. Nowadays, the classification of bread wheat in different quality categories is accomplished by expert assessors on the basis of the Italian Synthetic Index of Quality (ISQ) which is assigned to each wheat sample considering selected chemical and rheological properties. In many cases the procedure is not straightforward, making the class assignation uncertain, thus leading to the possibility of controversies. Furthermore, the determination of the ISQ parameters is time consuming and requires expensive laboratory analyses. In order to solve this issue, Partial Least Squares (PLS) calibration algorithm was applied to NIR spectra acquired on samples in three different physical forms to quantitatively predict the ISQ proprieties. Using different kind of PLS approaches (PLS1 and PLS2) the correlation between NIR spectra and ISQ parameters has been demonstrated. Promising results have been obtained for the determination of hectoliter weight values and the rheological parameter W.

Concerning the characterization by NIR of swine adipose tissue, the research was focused on the development of a fast and non-destructive method to classify two subcutaneous layers having different fatty acid compositions and destined to the manufacturing of different products. To this purpose, fat samples were analyzed by means of colorimetry, NIR spectroscopy and traditional destructive analysis such as Iodine Value determination (IV) and gas chromatographic measurements aimed to determine the amount of specific fatty acids as well as their overall amounts in terms of Saturated Fatty Acids (SFA) and Poly Unsaturated Fatty Acids (PUFA). A classification of the two layers was obtained by applying PLS-Discriminant Analysis (PLS-DA) to colorimetric and NIR data, also evaluating different spectral preprocessing methods and different variable selection algorithms. The classification results confirmed the possibility to recognize the two fat layers with efficiency equal to 78% for colorimetric data and 96% for NIR data. As for quantitative analyses, the calibration models showed high prediction performances for the evaluation of the degree of unsaturation represented by IV, linoleic acid, SFA and PUFA with correlation coefficients in prediction around 0.8; slight improvements have been obtained using a variable selection algorithm.

Keywords:
FT-NIR Spectroscopy
Chemometrics
Food
Wheat
Swine Fat

List of abbreviations:

a*: Chromaticity coordinate in CIE 1976 Lab color space;
b*: Chromaticity coordinate in CIE 1976 Lab color space;
BiPLS: iPLS calculated in reverse, or Backward mode;
B. U.: Brabender Units;
coif: Wavelet of the "coiflets" family;
CA: Classification Ability;
CAB: Classification Ability Basis;
CB: Contiguous Blocks cross validation method;
$CHCl_3$: Chloroform;
CIE: Commission Internationale de l'Eclairage;
C10: Capric acid i.e., decanoic acid;
C12: Lauric acid i.e., dodecanoic acid;
C14: Myristic acid i.e., tetradecanoic acid;
C16: Palmitic acid i.e., hexadecanoic acid;
C16:1: Palmitoleic acid i.e., 9*cis*-hexadecenoic acid;
C17: Margaric acid i.e., heptadecanoic acid;
C17:1: Heptadecenoic acid;
C18: Stearic acid i.e., octadecanoic acid;
C18:1: Oleic acid i.e., 9*cis*-octadecenoic acid;
C18:2: Linoleic acid i.e., 9*cis*12*cis*-octadecadienoic acid;
C18:3n-3: α Linolenic acid, ω3 i.e., 9*cis*12*cis*15*cis*-octadecatrienoic acid;
C18:3n-6: γ Linolenic acid, ω6 i.e., 6*cis*9*cis*12*cis*-octadecatrienoic acid;
C19: Nonadecanoic acid, internal standard in GC analyses as methyl ester;
C20: Arachidic acid i.e., eicosanoic acid;
C20:1: Eicosenoic acid;
C20:2: Eicosendioic acid;
C20:3: Eicosentrioic acid;
C20:4: Arachidonic acid i.e., all-*cis*-5,8,11,14-eicosatetraenoic acid;
CRA: Council for Research and experimentation in Agriculture (Consiglio per la Ricerca e la sperimentazione Agricola);
Chroma: Spectral saturation in CIE 1976 Lab color space;
Custom CV: Custom cross validation method;
CV: Cross Validation;
d1: First order derivative pretreatment;
d2: Second order derivative pretreatment;
db: Wavelet of the "daubechies" family;
det1: Linear detrend pretreatment;
det2: Quadratic detrend pretreatment;
D65: CIE standard illuminant;
EFF: Efficiency %;
E_{vib}: Potential energy of the system;
FA: Fatty Acids;
FAU: Flour for other uses (Farine Altri Usi);
FB: Wheat for biscuits (Frumento per Biscotti);
FF: Improver wheat (Frumento di Forza);
FiPLS: iPLS calculated in Forward mode;
FIR: Far InfraRed;
FN: Falling Number;

FOP: Fiber Optic Probe;
FP: Ordinary bread making wheat (Frumento Panificabile);
FPS: Superior bread making wheat (Frumento Panificabile Superiore);
FT-NIR: Fourier Transform-Near InfraRed;
GC: Gas Chromatography;
HSI: HyperSpectral Imaging;
Hue: Spectral color in CIE 1976 Lab color space;
I_2: Elemental Iodine;
I^-: Iodide ion solution;
In: Class corresponding to the inner layer of the fat samples;
InAs detectors: Indium-Arsenic detectors;
In_low: Samples of class *In* measured on the lower face of the fat sample;
InGaAs detectors: Indium-Gallium-Arsenic detectors;
In_up: Samples of class *In* measured on the upper face of the fat sample;
iPLS-DA: Interval Partial Least Squares-Discriminant Analysis;
iPLS: Interval Partial Least Squares;
ISQ: Italian Synthetic Index of Quality (Indice Sintetico di Qualità);
IS: Integrating Sphere;
IV: Iodine Value;
KI: Potassium Iodide;
KOH: Potassium Hydroxide;
L*: Brightness in CIE 1976 Lab color space;
Lab.1: Laboratory located in Department of Life Sciences of the University of Modena and Reggio Emilia;
Lab. 2: Laboratory located at CRA in S. Angelo Lodigiano
Lab.3: Laboratory located at the Granaria Association in Milan.
LED: Light Emitting Diodes;
LOO: Leave One Out cross validation method;
LV: Latent Variable;
m: Meancentering pretreatment;
MIR: Middle InfraRed;
MPA: Multi Purpose Analyzer;
MSC: Multiplicative Scatter Correction pretreatment;
MUFA: Mono Unsaturated Fatty Acids;
N: None pretreatment (raw spectra);
$Na_2S_2O_3$: Sodium Thiosulfate i.e., $S_2O_3^{2-}$ thiosulfate ion solution;
$Na_2S_4O_6$: Sodium Tetrathionate i.e., $S_4O_6^{2-}$ tetrathionate ion solution;
NIR: Near InfraRed;
Out: Class corresponding to the outer layer of the fat samples;
Out_low: Samples of class *Out* measured on the lower face of the fat sample;
Out_up: Samples of class *Out* measured on the upper face of the fat sample;
PA: Percentage of Assignation;
PbS: Lead Sulphide;
PC: Principal Component;
PCA: Principal Component Analysis;
Phl: Hectolitre weight;
P/L: Alveograph P/L ratio;
PLS: Partial Least Squares;
PLS-DA: Partial Least Squares-Discriminant Analysis;
PRESS: Predicted Residual Error Sum of Squares;

Prot_ss%: Dry matter protein content;
PUFA: Poly Unsaturated Fatty Acids;
Q-T^2: Q residuals versus Hotelling's T^2;
RD: Reduced Distance;
RG: Random Groups cross validation method;
RMSE Exp: Root Mean Square Error Experimental;
RMSEC: Root Mean Square Error in Calibration;
RMSECV: Root Mean Square Error in Cross Validation;
RMSEP: Root Mean Square Error in Prediction;
RT: Room Temperature;
S: Smoothing pretreatment;
SDS: Sodium Dodecyl Sulphate sedimentation value;
SENS: Sensitivity %;
SF: Score Function;
SFA: Saturated Fatty Acids;
SIMCA: Soft Independent Modeling of Class Analogy;
SNV: Standard Normal Variate pretreatment;
SPEC: Specificity %;
SSY: Sum of Squares Y explained variance;
Stab: Farinograph Stability;
sym: Wavelet of the "symlets" family;
TE: Thermo Electrically cooled;
TE-InGaAs detectors: Thermo Electrically cooled InGaAs light detectors;
TE-PbS photoconductors: Thermo Electrically cooled PbS light detectors;
TRACE™ GC Ultra apparatus: Finningan GC instrument;
Trilinolein: Glycerol tri C18:2 i.e., glycerol tri 9*cis*12*cis*-octadecadienoate;
Triolein: Glycerol tri C18:1 i.e., glycerol tri 9*cis*-octadecenoate;
Tristearin: Glycerol tri C18 i.e., glycerol tri octadecanoate;
TRN: Training set;
TST1: First test set;
TST2: Second test set;
UFM: Ultra Fast Module;
VB: Venetian Blinds cross validation method;
VIN: Contribution Variable Influence;
VIP: Variable Importance in Projection;
W: Alveograph W;
Wheat _F: Wheat as white flour;
Wheat _G: Wheat as grit;
Wheat _S: Wheat as wholemeal;
WPMAT: WPTER three dimensional array;
WPT: Wavelet Packet Transform;
WPTER: Wavelet Packet Transform for Efficient pattern Recognition.

To Me,
my Parents,
my Friends,
my Love,
my Passions,
my Teachers
and my Students.

To all of you, Thanks

TABLE OF CONTENTS

Chapter 1: INTRODUCTION

1.1 The concept of quality

The concept of "food quality" is twofold, in fact its first meaning is linked to *functional quality*, i.e., a desirable attribute of a product; while its second meaning refers to *conformance quality*, i.e., the production of a product that exactly meets the consumer's specification (Warris, 2000). The International Organization for Standardization (ISO) defines the notion of quality as "a set of characteristics able to satisfy the expressed or not expressed demand of the consumer" (Dell'Orto and Sgoifo Rossi, 2000).

In the context of food, quality represents a concept that is related to a large number of variables and parameters. Some of the quality parameters are objective and measurable, the ones deriving from chemical, physical and rheological analyses, while others are highly subjective and not measurable, since they are derived from different aspects of food perception, depending on historical, ethnical, and religious conditionings (Morrissey et al., 1998). The notion of quality can be also modified with the contemporary trend of consumer, in which the influence of the advertising plays a great role, especially for standardized products. The satisfaction of the foodstuff quality demand is extremely complex and related to a series of multi-factorial components, including different fields of studying such as health and safety standards (absence of residues and additives, pathogens and their metabolites), nutritional parameters (high digestibility, sufficient content of vitamins, essential fatty acids, antioxidants), technological and organoleptic factors. As a consequence, food quality is hard to define in a unique way and sometimes the standard definition results extremely variable in time and space. Hence, to efficiently reach a standard of quality several controls are required.

Food industry reached many different functional quality control tasks, such as keeping constant the flavours of food (as well as food estimation by the consumers), identifying sources of variability in food transformation processes that may lead to a change in quality, detecting adulteration in any food components and verifying raw materials in terms of geographical origin. In addition, to ensure consumers, many regulatory agencies, e.g., Food and Drug Administration in the United States or Protected Denomination Origin agencies in Europe, were born over the years. Most of them are interested in detecting health risks from possible food contamination as well as economic fraud and adulteration, in accordance with law and regulations.

Some of the needs and the problems that the food researcher and manufacturers have to face can be synthesized as follows:

- A global monitoring of changes in process parameters related to product quality from the starting material to the final product is often necessary in food transformation. As mentioned before, the target is maintaining the final food characteristic constant in terms of appearance and physico-chemical properties, i.e., texture, flavour, colour and shelf life. During the food production process, the unavoidable changes in the properties of raw materials have to be faced using a rational approach, in order to maintain the characteristics of the final products consistent and repeatable over time. For these reasons, food producers must have full knowledge of the chemical and physical transformations that occur to raw materials throughout the production process. In particular, the role of each ingredient and the possible synergy with the other ingredients during the manufacturing process should be known, to guarantee the rational control of the food processing. The issue to maintain the quality of the final product constant in time implies also that any variation in the used raw materials requires an appropriate and sudden action in the process conditions, able to compensate for this variation (Dell'Orto and Sgoifo Rossi, 2000).
- Periodic test of the properties of incoming raw materials. The aim is to be sure that starting materials meet certain minimum standards of quality if starting materials do not reach the standards required, the manufacturers can decide to reject them. Because of the variability of the raw materials during the storage, even a batch that was originally accepted might lead to changes in the properties of the final product. In this case, the analysis of the raw materials before use, allows to notice the change and to modify the conditions of the process to obtain even so a product with the desired properties.
- Development of instrumental systems able to replace the human sensory evaluation. The desirability of a food is determined by its interaction with the sensory organs of human beings (i.e., sight, taste, smell, touch and hearing), but unfortunately the individual perceptions of sensory attributes are often quite subjective. The use of statistical tools for dealing with sensory data can solve some problems linked to the human appreciation (Lea et al., 1997), but does not provide objective parameters such as chemical composition, safety and nutritional value of the food. Furthermore, the acquisition and processing of sensory data remains expensive, time consuming and, in most cases, the tests are not sufficiently objective. Sensory analysis is often used as the

ultimate test for the acceptance or rejection of a particular food product, considering consumers both trained or not. For these reasons, objective analytical tests that can be performed in a laboratory with standardized equipment and procedures are often preferred to check the properties of the food products that are directly related to sensory attributes. Many studies are currently in progress to correlate the sensory attributes (such as chewiness, tenderness or viscosity) to properties that can be measured using objective analysis techniques (Foca et al., 2011; Jha and Matsuoka, 2000).

- Effective detection of the adulteration, the contamination and the replacement of a product. Since the price of some foods is linked to the quality of the ingredients they contain, it is not uncommon that food producers make false statements about the authenticity of their products in order to make more money. Hence, it is important, sometimes mandatory, the use of analytical techniques able to verify the authenticity of some food components, in a manner that the consumers are not victims of economic fraud and that the fair competition among producers is protected (Wilson and Defernez, 1997; Downey, 1995).

1.2 The role of NIR spectroscopy in food analysis

In an efficient food factory, frequent and systematic analyses on the entire productive process are needed. These analyses have to be fast and cheap and they have also be able to detect unwanted variations in the production process as soon as possible. From the analytical point of view, complex matrices such as foodstuffs require a preliminary step for the separation of the different chemical components, in a way to eliminate the interfering substances, before the out-and-out analysis. For these reasons, although the traditional determinations are generally very accurate, in most of the cases they are expensive, time consuming and they need to employ qualified personnel. A different analytical approach is represented by the investigation of the food matrix as a whole, without looking for its distinct components. This approach has also the advantage to consider the physico-chemical interactions existing among all the sample components, the so-called “matrix effect”.

The spectroscopic techniques are particularly suitable to reach this aim; in particular, among all the spectroscopic techniques which can be employed for the characterisation of a complex matrix, Near InfraRed (NIR) spectroscopy surely plays a primary role, for a number of reasons.

First of all, NIR spectroscopy is a vibrational spectroscopy suitable to investigate organic matrices, such as foodstuffs. When a complex matrix is analyzed without making a preliminary preparation of the sample, the NIR spectral profile obtained is a sort of fingerprint of the sample. This implies that the NIR spectrum is certainly difficult to interpret in classical terms, but it is not necessary to interpret it punctually, since considering it as a whole it is possible to obtain unique information on the sample analyzed. In addition, NIR spectroscopy is advantageous in the food industry as it provides the analytical response in very short times and it is a non-destructive technique. For these reasons, this analysis could be used as on-line or in-line process monitoring by an appropriate automation of the quality control system. Finally, NIR spectroscopy is particularly easy to use, even by non-experts, and it is relatively inexpensive; these factors are the causes of the current wide diffusion of this technique in the laboratories for quality control.

The InfraRed (IR) region of the electromagnetic spectrum was discovered in 1800 as the "*invisible radiant heat"* of the sun and it was described as the radiation able to heat bodies (Herschel, 1800). Subsequently, the interaction between matter and IR light was investigated with the birth of infrared spectroscopy. For analytical purposes, the IR spectral region was then divided in three sub-regions: the NIR region, from 0.78 to 2.5 μm (i.e., 13000-4000 cm^{-1}), the Far InfraRed (FIR) region, from 50 to 100 μm (i.e., 200-10 cm^{-1}) and the Middle InfraRed (MIR) region, from 2.5 to 50 μm (i.e., 4000-200 cm^{-1}).

From the historical point of view, NIR spectroscopy had a slow start. While in the MIR region the pure organic compounds identification is quite easy, in the NIR region the spectrum of a single compound is composed by many overlapping peaks related to numerous overtones and combination bands of the vibrational transitions, making spectral interpretation rather hard. Moreover, the absorption bands in the NIR range are very weak (by two or three order of magnitudes lower than in the MIR range), and because of the overall complexity, baselines are hard to define. Therefore, to extract useful information from the whole NIR spectrum, bearing information on the chemical composition and physical state of the whole sample, multivariate data analysis, i.e., chemometrics, is needed.

The use of NIR technology in transformation processes and food quality determination took a leap in the second part of 20^{th} century, mainly for the development of the spectroscopic instrumentation and for the exponential growth of the computer performances, that have improved the speed of data elaboration (Alander et al., 2013; Huang et al., 2008; Cen and He, 2007). In this respect, credit has largely to be given to Karl Norris, who recognised the

potential of NIR spectroscopy from the early stages of its development and applied it to the agricultural science field. Since the late 50s only a few works regarding NIR spectroscopy, that was used to solve calibration issues about agricultural products, have been published by Norris and other researchers (Birth and Norris, 1958; Ben-Gera and Norris (1968a, 1968b)). The 70s were the years when NIR spectroscopy started to show all its resources, since it was applied with success for the characterization of food samples.

Nowadays, the NIR region is preferred for calibration purposes with respect to the other IR region, because of its richness in absorption bands of different intensity bringing the same chemical information. In fact, even if the FIR and MIR spectra can be also interpreted after visual inspection by expert personnel, when we are dealing with a complex matrix the use of chemometrics becomes mandatory (Wilson, 1990; Wilson and Kemsley, 1996; Wilson and Tapp, 1999; Downey, 1998).

1.3 The importance of chemometrics: a multivariate approach to data analysis

Chemometrics is a very useful tool, sometimes mandatory, to examine of the complex data coming from complex matrices. According to the definition of Workman (2002) and Workman et al., (1996), chemometrics can generally be described as the application of mathematical and statistical methods to improve chemical measurement processes and to extract more useful information from chemical and physical experimental data.

Almost all real systems or issues cannot be described by means of a clearly defined theory. In general, theories give a background of knowledge that helps to face the problem, but does not allow a specific problem resolution. In fact, most of the several variables potentially implied are not controllable with the desired precision. For many variables, it is not possible to exactly know the importance for the considered samples and the amount of noise that can confuse the real effects of the considered variables. In addition, it is generally not possible to know the presence of synergic effects among the different variables or their correlations. In many cases, the systems do not show linear relationships between selected variables, as a consequence, the problem complexity cannot be initially described with linear models.

The chemometric methods are able to handle the data complexity extracting all the useful information contained in the data and examining also the interactions among variables by means of multivariate methods. Essentially, the chemometric techniques allow data

exploration, with the aim to investigate the data pattern and structure, the relations and correlations among data (samples and variables). Moreover multivariate data exploration allows the graphical representation of data. In most of the cases the chemometric approaches are able to separate the relevant and useful information of the data from not pertinent information–good quality information, but not directly relevant for the problem at hand–and from the presence of experimental noise (Figure. 1.3.1).

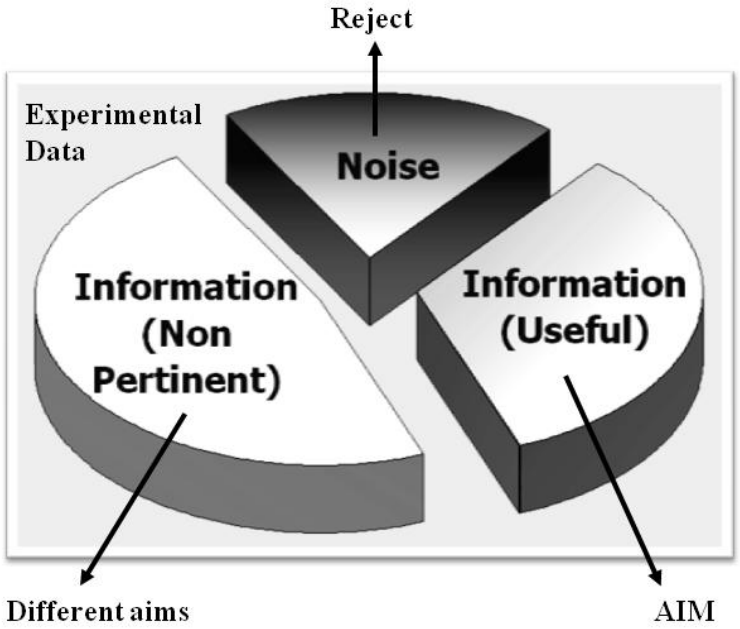


Figure 1.3.1. Separation of the different sources of information in the experimental data by means of explorative analysis.

The most rational approach to factually face the real problems consists in using all the information available from the previous experience and to optimise the procedures for the development of new strategies both with the lowest times and costs and with the highest quality as possible. From these needs, the chemometric methods were born, by using knowledge from different fields of science such as statistics, mathematics, computer science, analytical chemistry and other experimental sciences.

Chemometrics is a group of multivariate statistical methods able to correlate quality parameters or physical properties (e.g., descrete experimental variables) to analytical instrumental data registered on food products (e.g., NIR spectra) (Miller and Miller, 2000; Beebe et al., 1998). Once the data patterns are correctly modelled, it is possible to apply the obtained model to future data in order to predict the same quality parameters. The result of the chemometric approach is an efficiency gain in assigning product quality. Of course, this

approach needs the use of an appropriate analytical instruments (e.g., a NIR spectrophotometer or other instruments) and software able to interpret the data patterns. Regardless of the type of data collected, chemometric software is able to recognize the patterns of data in any type of multidimensional analytical data. Chemometrics is also employed to speed up the development of analytical methods and to render the use of statistical models for data analysis a common practice.

In other words, the chemometric methods often seem to be the only way to deal with highly complex problems, for which theoretical well founded strategies are not available. The great development of the computer world (in terms of computational speed and data handling and storing) and its accessibility at low costs have permitted the development of chemometrics, that can offer to the researchers a wide gamut of potential methods and solutions, that can be tested and checked in short times.

1.4 Wheat and swine adipose tissue: two specific case studies

In this Ph.D. thesis, attention has been focused on the development of new applications of NIR spectroscopy coupled with chemometric methods for the extraction of useful information, in a manner to provide to the food industry further tools for improving the quality control.

In particular, two main food matrices have been characterized by means of this approach: wheat and swine adipose tissue. Both matrices have a deep relevance from the economical point of view in the north of Italy, since these foodstuff productions involve millions of euro and employ thousands of workers (FAO Food Outlook Report, 2012; Rabobank, 2012; USDA Foreing Agricultural Services, 2012).

More in detail, as for wheat, the possibility to use NIR spectroscopy for the prediction of quality of raw matters has been investigated. Nowadays, the classification of bread wheat in different quality categories is accomplished by expert assessors on the basis of the Italian Synthetic Index of Quality (ISQ) which is assigned to each wheat sample considering selected chemical and rheological properties. In many cases the procedure is not straightforward, making the class assignation uncertain, thus leading to the possibility of controversies (Cocchi et al., 2005; Foca et al., 2007). Furthermore, the determination of the ISQ parameters is time consuming and requires expensive laboratory analyses. In order to solve this issue, multivariate calibration was applied to NIR spectra acquired on samples in three different physical forms to quantitatively predict the ISQ proprieties.

Concerning the characterization of the swine adipose tissue, the research was focused on the development of a fast and non-destructive procedure, based on NIR spectroscopy and colorimetry, to classify two subcutaneous layers having different fatty acid compositions and destined to the manufacturing of different products. In addition, a further work has been done to verify the possibility to replace the traditional destructive analysis, i.e., Iodine Value and gas chromatographic determinations, aimed to measure, respectively, the degree of unsaturation of the pig fat and the amount of specific fatty acids, with a NIR based non-destructive method.

Finally, it is important to draw some considerations concerning the outcomes of the proposed methods in the field of applied analytical chemistry. Actually, this kind of approach is very important in the backstage of the analytical process. In fact, it needs an heavy and time-consuming preliminary work, in order to prepare the method and to fully validate it, but then it can furnish innumerable advantages to the final user. The end-user, in fact, can routinely use the method developed by chemometricians as a *black box*, in which it is sufficient to put the initial information to gain the analytical response of interest. In this context, the last goal of this doctoral thesis work is to develop and to enhance new methods, able to simplify the analytical controls for the final analyst, consequently intensifying the number of the controls and to improving their quality.

1.5 References

Alander, J.T., Bochko, V., Martinkauppi, B., Saranwong, S., and Mantere, T., (2013). A review of optical nondestructive visual and near infrared methods for food quality and safety. *Internat. J. Spectrosc.* Vol 2013, Hindawi Publishing Corporation.

Beebe, K.R., Pell, R.J., Seasholtz, M.B., (1998). Chemometrics - A practical guide. Ed. Wiley & Sons Ltd, New York, US.

Ben-Gera, I., Norris, K.H., (1968a). Determination of moisture content in soybeans by direct spectrophotometry. *Israeli Journal Agr. Res.* 18, 124-132.

Ben-Gera, I., Norris, K.H., (1968b). Direct spectrophotometric determination of fat and moisture in meat product . *J. Food Sci.* 7, 240.

Birth, G.S., and Norris, K.H., (1958). An instruments for using light transmittance for Nondestructive measurements on fruit maturity. *Food Tech.* 12, 592.

Cen, H. and He, Y., (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends Food Sci. Tech.* 18, 72-83.

Cocchi, M., Corbellini, M., Foca, G., Caramanico, R., Lucisano, M., Ambrogina Pagani, M., Tassi, L., Ulrici, A., (2005). Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra. *Anal. Chim. Acta* 544, 100-107.

Dell'Orto, V., Sgoifo Rossi, C.A., (2000). Aspetti nutrizionali per la produzione di carne bovina di qualità. *L'Informatore Agrario.* 14, 45-56.

Downey, G., (1998). Food and food ingredient authentication by mid-infrared spectroscopy and chemometrics. *Trends Anal. Chem.* 17(7), 418-424.

Downey G. (1995). Food quality and autenticity measurements. *Farm & Food*, Jul-Sept, 21-24.

FAO Food Outlook Report (2012), Global Market Analysis, (November 2012) 1-125, www.fao.org

Foca, G., Ulrici, A., Corbellini, M., Pagani, M.A., Lucisano, M., Franchini, G.C., Tassi, L., (2007). Reproducibility of the italian ISQ method for quality classification of bread wheats: An evaluation by expert assessors. *J. Sci. Food Agric.* 87 (5), 839-846.

Foca, G., Masino, F., Antonelli, A., Ulrici, A., (2011). Prediction of compositional and sensory characteristics using RGB digital images and multivariate calibration techniques. *Anal. Chim. Acta* 706, 238-245.

Herschel, W., (1800). Experiments on the refrangibility of the invisible rays of the sun. *Phil. Trans. R. Soc. Lond.* 90, (XIV), 284-293.

Huang, H., Yu, H., Xu, H. Ying, Y., (2008). Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review. *J. Food Eng.* 87, 303-313.

Jha, S.N., Matsuoka, T., (2000). Non-destructive techniques for quality evaluation of intact fruits and vegetables. *Food Sci. Tech. Res.* 6 (4), 248-251.

Lea, P., Naes, T., Rodbotten, M., (1997). Analysis of variance for sensory data. Ed. Wiley & Sons Ltd, New York, US.

Miller, J.N., Miller, J.C., (2000). Statistics and chemometrics. 4th Ed. Prentice Hall.

Morrissey, P.A., Sheehy, P.J.A., Galvin, K., Kerry, J.P., and Buckley, D.J., (1998). Lipid stability in meat and meat products. *Meat Sci.* 49, S73-S86.

Rabobank, (2012). Industry Note #328 European Flour Milling Industry. 1-7

USDA Foreing Agricultural Services (2012). Italian Grain and Feed Report, *Global Agricultural Information Netwotk (GAIN) Report*.

Warris, P.D., (2000). Meat science- An introductory text. CABI Publishing.

Wilson, R., Defernez, M., (1997). Foiling the food fraudsters. *Chem. Rev.* Sept, 53-58.

Wilson, R.H., Kemsley, E.K., (1996). Test of truth. *Lab. Tech. Internat.* 81-83.

Wilson, R.H., (1990). Food analysis with mid-infrared spectroscopy. *Spectrosc. World* 2(1), 40.

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

Workman, J. Jr., (2002). The state of multivariate thinking for scientists in industry: 1980-2000. *Chemom. Intell. Lab. Syst.* 60, 13-23.

Workman, J.J., Mobley, P.R., Kowalski, B.R., Bro, R., (1996). Review of chemometrics applied to spectroscopy: 1985-95, Part I, *Appl. Spectrosc. Rev.* 31, 73-124.

Chapter 2: NEAR INFRARED SPECTROSCOPY

2.1 Some theory

The electromagnetic radiation is considered as energy which exhibits wave-like behavior as it travels through space. Electromagnetic radiation can interact with matter. In fact, it can be either absorbed or transmitted by a substance, depending upon its frequency and on the chemical composition on the matter sample. By absorbing radiation, a molecule can temporary increase its own energy and let happen a quantum transition from one low level energy state ($E_{initial}$) to an high level energy state ($E_{final} > E_{initial}$). The transition results possible only if the energy gap is in accord with Planck's law:

$$E_{final} - E_{initial} = \Delta E = h\,v = hc/\lambda \qquad (2.1)$$

If a transition exists, it means that the frequency of the incident radiation is in accord with Planck's constant, and the radiation can be absorbed. Otherwise, the radiation will be transmitted. The type of absorption spectroscopy derives from the type of transition involved, and thus from the wavelength range of the electromagnetic radiation absorbed.

In the case of heteronuclear molecules, the radiation of a specific region of the electromagnetic spectrum can interact with molecule vibrations: the infrared radiation absorbed by the molecule allows the transition from one vibrational energy level to another, and then the related spectroscopic technique is known as infrared (IR) spectroscopy (Cozzi et al., 1997; Coates, 2000).

IR spectroscopy is a technique based on the absorption, by molecules or compounds, of radiations with wavelengths between 800 nm (the limit of the visible region) and $2 \cdot 10^6$ nm. This wide spectral range is split in three main regions (Figure 2.1.1):

- Near IR (NIR): between 800 and 2500 nm (about 12500-4000 cm^{-1});
- Middle IR (MIR): between 2500 and 25000 nm (4000-400 cm^{-1});
- Far IR (FIR): between 25000 and $2 \cdot 10^6$ nm (400-5 cm^{-1})

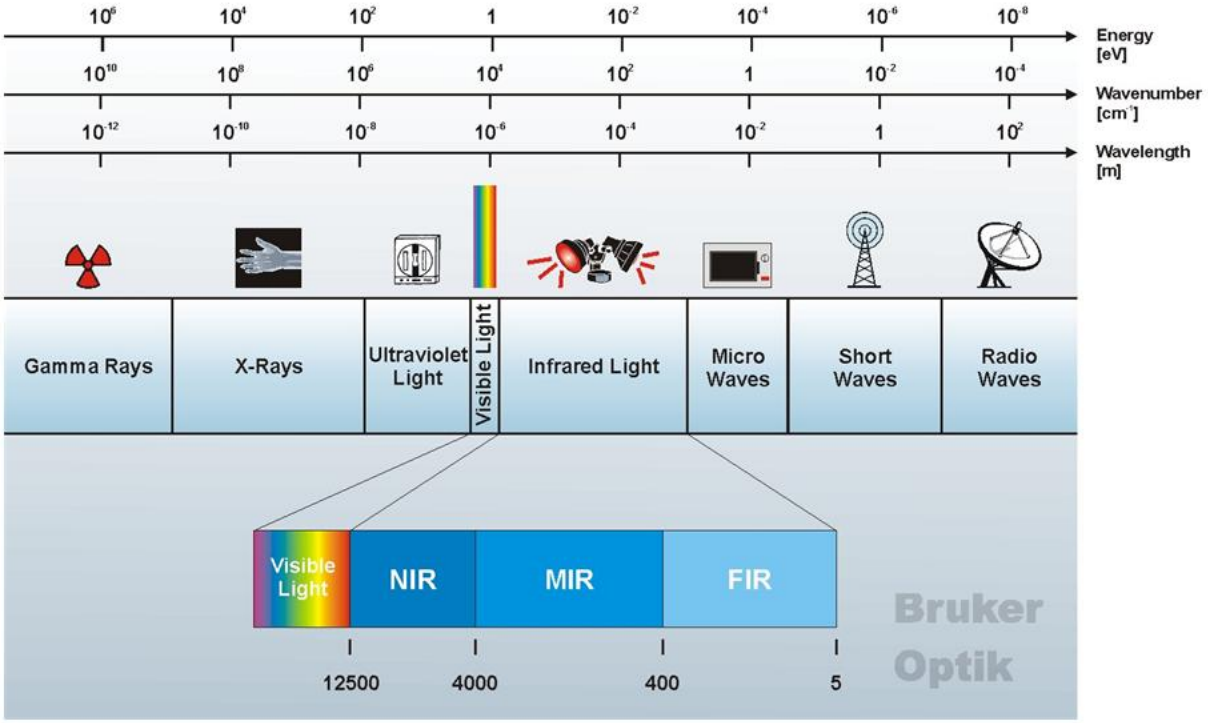


Figure 2.1.1. The electromagnetic spectrum (Bruker Optics).

Each region is used in different analytical fields. The MIR region permits to identify the typical vibrations of functional groups of organic molecules and furnishes information about the structure of the analysed compounds. The NIR region instead presents complex groups of overlapped bands that can be linked, with some difficulties, to few functional groups, while the FIR region is more interesting in order to characterize inorganic and organo-metallic compounds.

Considering the molecule from a dynamic point of view, the IR radiation is able to amplify the natural periodic variations of the inter-atomic distances and the bond angles in the molecules. In undergoing such a transition, the molecule gains vibrational energy, and this is manifested by an increase in the amplitude of the vibration. The frequency of light required to cause a transition for a particular vibration is equal to the frequency of that vibration, so that we may measure the vibrational frequencies by measuring the frequencies of light which are absorbed by the molecule.When a molecule absorbs an infrared radiation a sharp change of the dipole moment does happen, as a result of its movement of vibration or rotation. If the interaction takes place, then the electric field associated with the light radiation can interact with the electric field originating from the fluctuation of the dipole moment of the molecule. If the frequency of the light radiation equals exactly the frequency of natural vibration of the molecule, a transfer of energy occurs. This gives rise to a change of amplitude of the

molecular vibration and, as a result, to the absorption of the radiation. In the case of homonuclear molecules (e.g., H_2 N_2, O_2) the dipole moment has no alteration during rotation or vibration, and, as a consequence, these molecules do not record absorption of radiation in the infrared spectrum.

For a real diatomic molecule, the variation of the inter-atomic distances as a function of the potential energy of the system (E_{vib}), is presented in Figure 2.1.2.

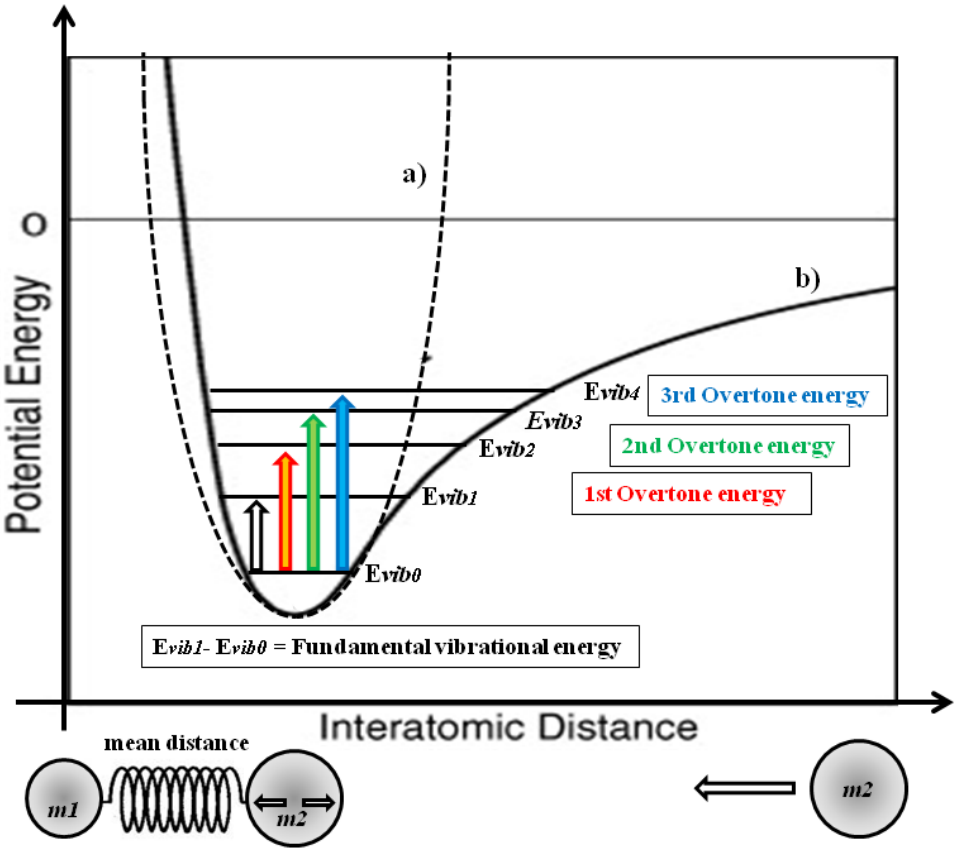


Figure 2.1.2. The variation of the inter-atomic distance as a function of potential energy in a diatomic molecule. The behavior of the classic harmonic oscillator is represented with the dashed line, while the behavior of the unharmonic oscillator is represented with the solid line.

A diatomic molecule can be represented as a system in which two masses are connected by a spring; this system is naturally in oscillation, with a frequency that depends on the extent of the masses and the force constant of the spring. The model described is called the *classical harmonic oscillator*. At the equilibrium length, the spring has lowest potential energy (e.g., zero corresponding to higher degree of spring-bond stability). As the spring is stretched or compressed, the potential energy increases following a parabolic curve (the dashed line in Fig.

2.1.2), that is called the *harmonic potential*. The vibrational frequency ν is related to the force constant *k* and mass *m* by the equation 2.2:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \qquad (2.2)$$

Therefore, a larger force constant, i.e., a stronger spring, results in a higher frequency, while a smaller force constant in a lower frequency. In the same manner, a larger mass results in a lower frequency, but the potential energy curve does not change. The molecule vibrates with a total energy equal to the potential energy at the stretched or compressed position (eq. 2.3):

$$E = \frac{1}{2}kx_{\max}^2 \qquad (2.3)$$

The model therefore predicts that the molecule can vibrate at any total energy. While this is true for a system composed by two real masses and a spring, it is incorrect for a molecule. A more accurate quantum mechanics model has to be used to predict the observed behavior of molecules. In fact, a molecule may vibrate only at energy levels which fit the formula in equation 2.4:

$$E_n = (n + \frac{1}{2})h\nu \qquad (2.4)$$

where *n* (*n* = 0, 1, 2, 3…) is the vibrational quantum number and *h* is the Planck's constant. The energy is said to be quantized. In the quantum mechanical model, a molecule may only absorb radiation of an energy equal to the difference between two energy levels. In addition, the selection rule Δn = ±1, states that for a harmonic oscillator these transitions can only occur from one level to the next higher. Hence, a molecule can absorb radiation only with energy equal to *h*ν; the infrared spectrum of such a molecule should have a single peak at the frequency corresponding to that energy.

A real spectrum, however, is more complicated than that was explained above. Firstly, a real molecule is not a harmonic oscillator, in fact, when atoms get closer to one another, they repel more strongly than a spring and when they are pulled apart far enough, the chemical bond breaks. This behavior can be described with an *anharmonic potential* (the solid line in Fig. 2.1.2). In this more realistic model, the energy levels are almost equally spaced only in the parabolic area of the harmonic potential. This fact has an important consequence: the selection rule that allows only the transitions between two adjacent levels is not rigorously true, in other words, also the transitions with Δn = ±2 and, less frequently, Δn = ±3 may

occur. These transitions, called overtones, appear in the IR spectrum at a little less than twice or three times the frequency of the fundamental absorption band and are less intense. Therefore, the fundamental vibrations of a molecule are located in the MIR region, while the overtone vibrations in the NIR region. In the NIR region also other types of absorption bands (called combination bands) are frequently observed; they are due to the interactions among the different fundamental frequencies of vibration of polyatomic molecules.

The spectral bands in the NIR region corresponding to overtones and combinations of the fundamental vibrational transitions, mainly ascribable to the presence of CH, NH, OH, and SH groups (Bokobza, 1998), are presented in Figure 2.1.3.

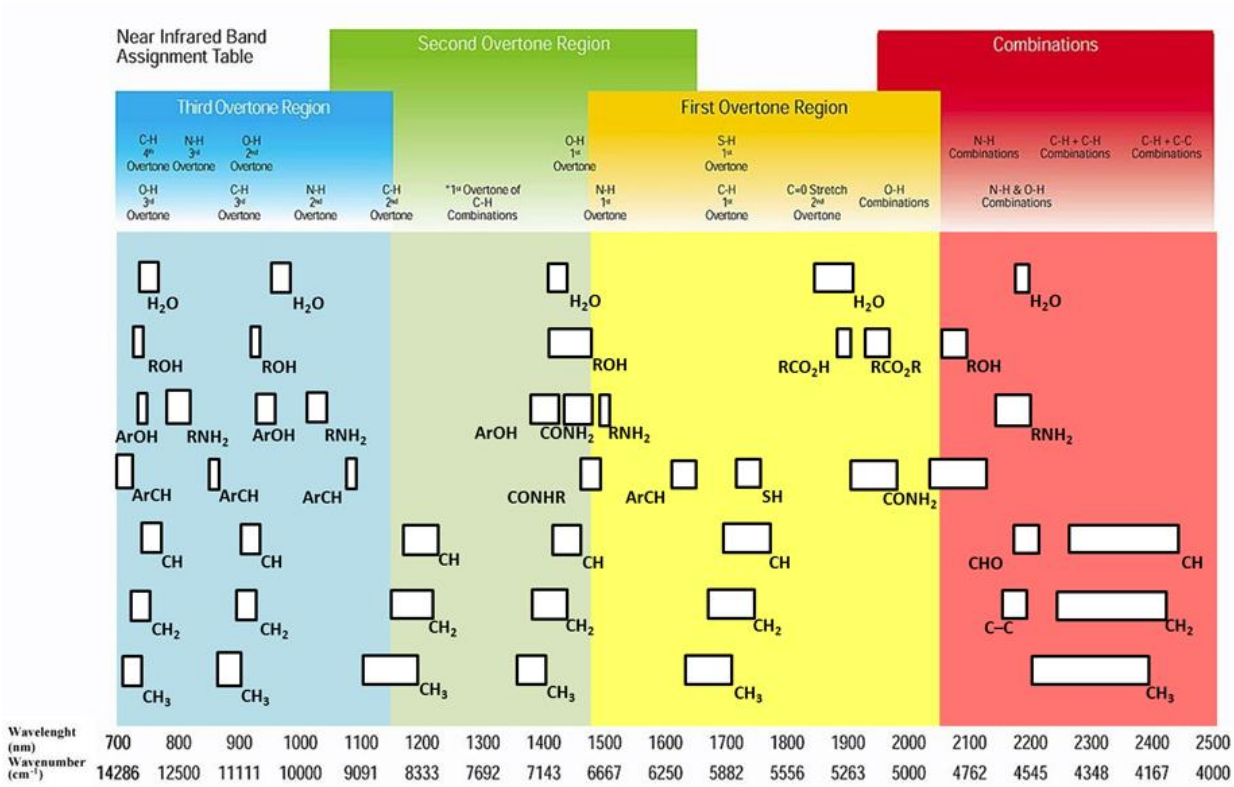


Figure 2.1.3. NIR bands assignment table (Bruker Optics).

NIR offers a number of advantages for quantitative analysis although its use in structural studies is limited. Since the NIR spectrum tends to consist of broad, featureless bands, the analytical potential of the NIR was largely overlooked until the pioneering work of Norris (Ben-Gera and Norris, 1968). Norris and co-workers showed that the spectral information in the NIR region could be extracted from the complex overlapping features present in the spectrum if statistical principles were used to evaluate the spectra. This observation has led to

the birth of powerful multivariate calibration strategies that made up a new field of chemometrics.

2.2 Basic instrument configuration

Instrumental setting for NIR spectroscopy

The essential scheme of a spectrophotometer is reported in Figure 2.2.1.

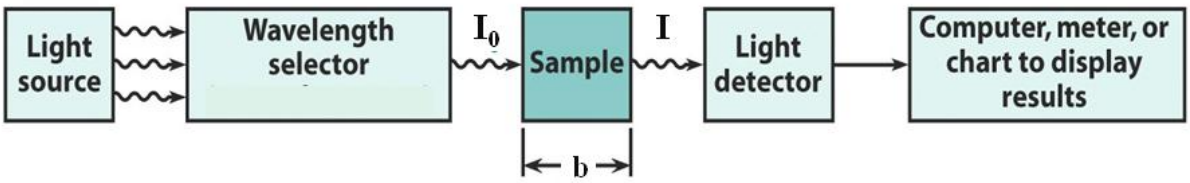


Figure 2.2.1. Basic spectrophotometer scheme.

Light source

The most common used light source in NIR spectroscopy consists in a tungsten-halogen lamp with quartz windows, able to produce a continuous spectrum in the region 320-2500 nm. Other, less common, light sources used are called LEDs (Light Emitting Diodes).

Wavelength selector

The NIR instrumentation can be classified depending on the wavelength selector used or, to be more precise, on the optical bench:

i. dispersive spectrophotometers, that contain a diffraction grating or prism or monochromator as wavelength selector to disperse the input beam, primarily used for qualitative works;

ii. Fourier Transform (FT) spectrophotometers, where an interferometer replaces the dispersive device (Rubinson and Rubinson, 2002).

Nowadays, the old dispersive spectrophotometers have been largely replaced by novel FT instruments. In fact there are some fundamental advantages in the use of FT-NIR instruments: high speed, reliability, convenience and their higher signal-to-noise ratio, necessary in a region with low amplitude of signal and high noise like NIR. Therefore, it may be worthwhile to bring some further details on the working principles of FT instruments.

The Michelson interferometer is composed by different parts (Figure 2.2.2). In this apparatus, the light emitted by the source passes through a beamsplitter, which sends the light

in two directions at right angles. One beam goes to a stationary mirror, the other goes to a moving mirror; then both the beams are reflected back to the beamsplitter, where they recombine. The motion of the moving mirror makes the total path length variable versus that taken by the stationary-mirror beam. The difference between the paths of the two beams is called delay (δ). Consequently, the difference in path lengths creates periodic partial interferences, total constructive interference (if $\delta = k\lambda$, where λ is each single wavelength emitted by the source) and total destructive interference (if $\delta = \frac{1}{2} k\lambda$). Ultimately, the interferometer creates a tuning of the radiation intensity that depends on the moving mirror speed: this is called *interferogram*. The radiation resulting from the interferometer passes through the sample, which absorbs all the different wavelengths characteristic of its spectrum, then subtracting specific wavelengths from the interferogram.

The detector reports the variation of the transmitted and/or absorbed radiation as function of the delay for all wavelengths simultaneously, hence the obtained signal is expressed in the time domain. A laser beam is superimposed to provide a reference for the instrument operation.

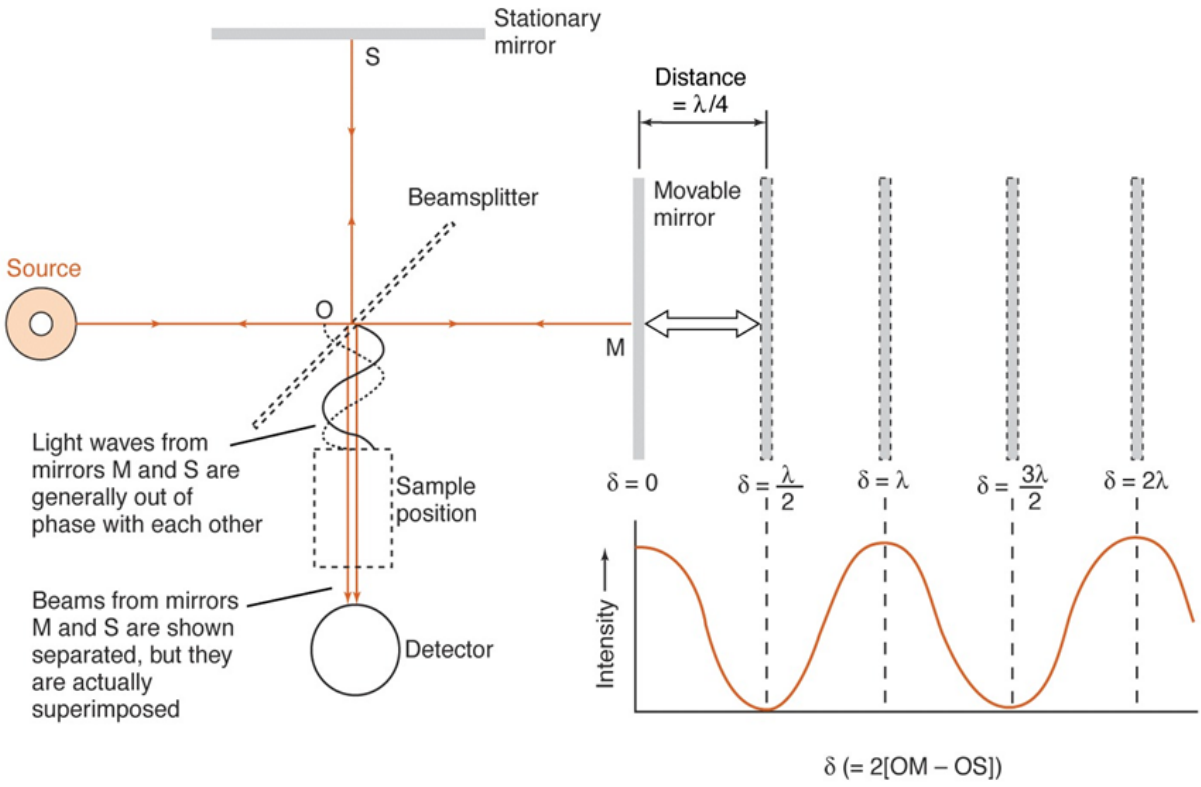


Figure 2.2.2. Schematic representation of the Michelson interferometer running.

The Fourier Transform mathematical function allows to decompose the interferogram into its composing λ and this means converting the signal from the time domain to the frequency domain, that is easier to understand in terms of chemical absorption of specific functional groups.

Sample cell

In NIR spectroscopy exist several methods to acquire the spectrum, and the sample cell used depend upon the sampling procedure chosen for the problem at hand. For this reason, a deepen explanation on this topic is reported in paragraph 2.3.

Light detector

The detectors measure the intensity of the radiation after its interaction with the sample and generally, in the NIR region, they consist in semiconductor-based devices. The most common used detectors are composed by lead sulphide (PbS) detectors. In particular, the thermo electrically cooled (TE)-PbS photoconductors are able to collect radiation into the range 780-2780 nm (12800-3600 cm^{-1}). For transmission measurements on solids at Room Temperature (RT), InAs and InGaAs detectors, able to detect light in the range 780-1725 nm (12800-5800 cm^{-1}), are used (Workman and Burns, 2001). When thermo electrically cooled, the (TE)-InGaAs detectors can reveal the infrared radiation over the whole NIR range 780-2500 nm (12800-4000 cm^{-1}). Other particular detectors used in NIR spectroscopy are the Si Diodes able to detect light in range 650-1100 nm (15500-9000 cm^{-1}) (MPA user manual).

The computer at the end of the scheme in Figure 2.2.1 permits to amplify, to convert and to record the analytical signal.

2.3 How to analyze a sample by NIR spectroscopy

Normally NIR instruments can record spectra on solid, liquid or gaseous samples. Since solid matrices, i.e. wheat and pig fat, have been employed in this work, in this section only the solid samples handling has been presented.

When a beam of NIR radiation hits the sample various phenomena can be observed, i.e. transmission, scattering and diffuse reflectance, depending on the angle of incidence of NIR light and on the characteristics of the sample (Figure 2.3.1). If a surface diffusely reflects without penetration of the light into the sample, like specular reflectance, no absorption takes

place. In the NIR region, scattering is so high that transmittance through 1 cm of most sample is negligible. Such a situation is called diffuse reflectance, because most of the incident radiation is reflected (Olinger et al., 2001). In typical NIR diffuse reflectance experiment a powdered sample is packed into a sample cell and covered with a quartz window, that is transparent up to about 3000 nm (Figure 2.3.2).

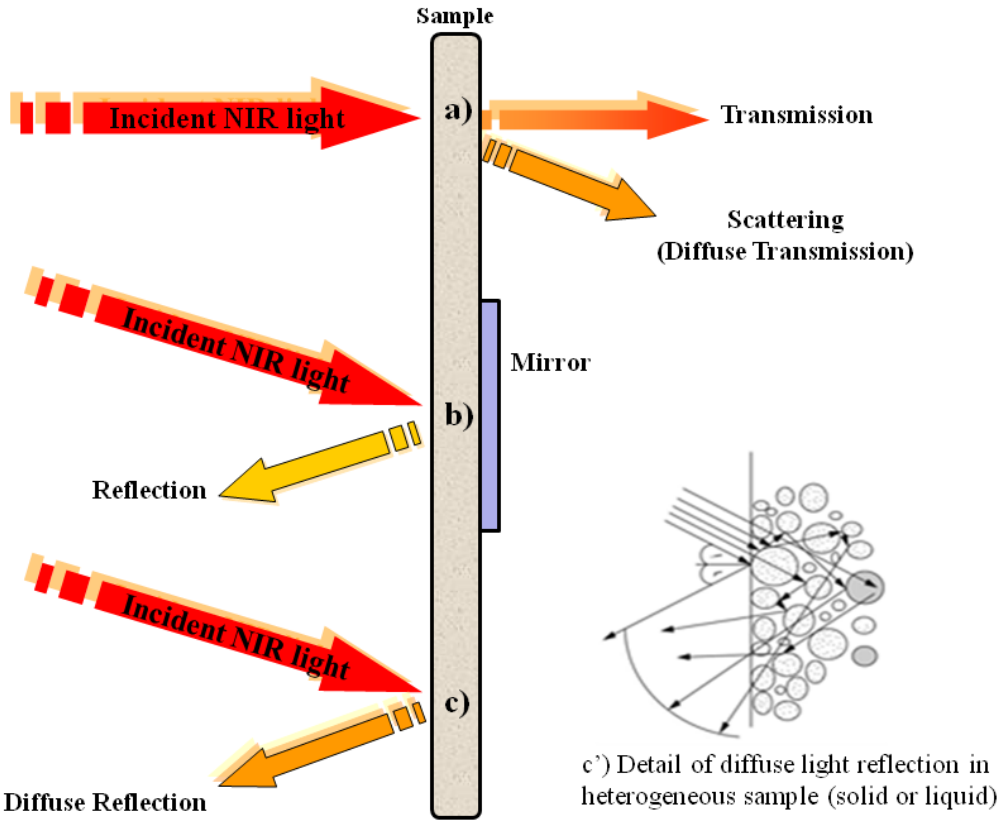


Figure 2.3.1. Different kinds of interaction between NIR light and sample.

Figure 2.3.2. Different sample cells with quartz window for NIR diffuse reflectance spectroscopy.

The sample cell is then placed into the instrument where it is illuminated with NIR radiation, and the reflected beam can be measured by using a set of detectors placed at 45° with respect to the incident beam, or by a single detector onto which the beam is focused by means of an integrating sphere (IS). The IS is a hollow sphere whose interior is coated with a highly reflective material, such as barium sulphate, having little holes to allow to the radiation to enter and exit. The IS is then used to analyze the samples which give diffuse reflectance or to render uniform the intensity of a light source over all positions within its circular aperture.

The reflectance measurements penetrate only from 1 to 4 mm of the front surface of ground samples. This small penetration (with respect to the sample thickness) of energy into a sample brings about greater variation with respect to transmittance techniques, in particular, when measuring nonhomogeneous samples. In transmittance measurements the entire pathlength of a sample is integrated into a spectral measurement, thereby reducing errors due to nonhomogeneity of samples. Transmittance techniques are most useful for measuring large particles (Nielsen et al., 2003), while for fine particles, the front surface scatter brings about a loss of energy transmitted through a sample with a net effect being a decrease in the signal-to-noise ratio of the instrument. In transmittance, higher frequency energy is most commonly used due to its greater depth of penetration into the sample. The higher frequency energy (800-1400 nm) is more susceptible to front surface scattering than lower frequency energy.

Among the several methods that can be used in NIR spectroscopy to investigate the sample, the Fiber Optic Probe (FOP) merits a special mention, since it is compact and rapid,

so it is suitable to be used for on-line quality control (Berntsson et al., 2001). It has the advantage that the measurement is very fast (one spectrum per second), so that the NIR instrument integrated in the production line would ensure that food products are processed efficiently, with less variation and waste.

A FOP essentially consists in a flexible cable containing tens of optic fibers made of polymeric material or silica glass. Each single transparent optic fiber is able to transport the light, between the two end of the fiber, by means of the physical principle called total internal reflection (Figure 2.3.3). When a beam passes from a more dense (the core) to a less dense (the cladding) medium a reflection occurs. The fraction of the incident beam that is reflected increases as the angle of incidence becomes larger; beyond a certain critical angle the reflection is complete (Coates and Sanders, 2000).

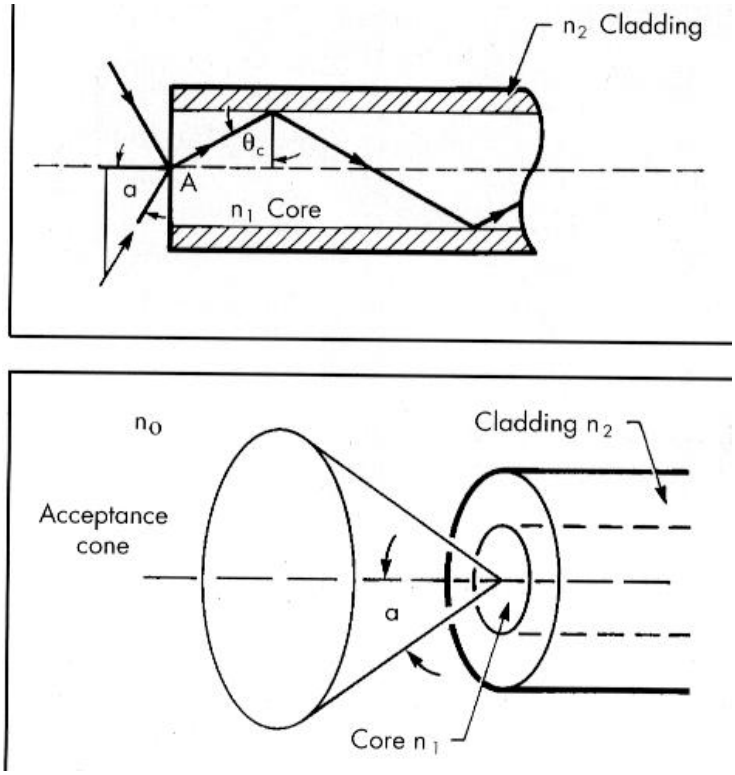


Figure 2.3.3. Total reflection in an optic fiber.

The flexibility of the fiber optic cable and the hand–held probes, together with the fact that light transfer is not affected by electric noise, permits the measurements of remote samples or even the direct measurement of stored raw materials, as shown in Figure 2.3.4.

Figure 2.3.4. Example of use of FOP for raw materials quality control.

2.4 The Bruker Optics Multi Purpose Analyser

In this thesis work, a Bruker Optics Multi Purpose Analyser (MPA) FT-NIR spectrophotometer was used. In Figure 2.4.1 the instrument equipped with all the available measurement modules is represented. (MPA brochure).

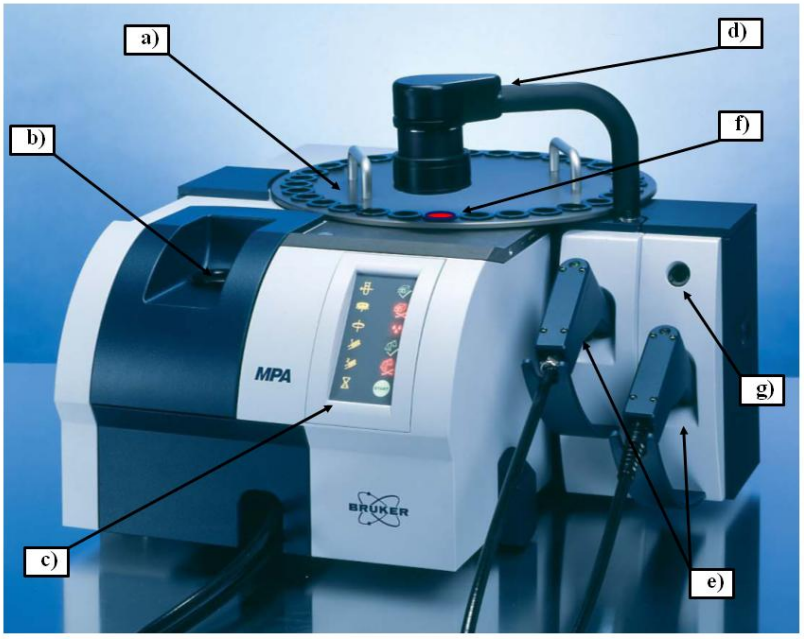


Figure 2.4.1. The MPA FT-NIR Bruker Optics spectrophotometer.

The different parts of the instrument, indicated in Figure 2.4.1 by letters, correspond to the following components:

a) sample changer with 30-position sample wheel;
b) sample compartment to place cuvettes or vials for temperature controlled transmission measurements. The detector used for these measurements is (TE)-InGaAs;
c) instrument display;
d) transmission unit to perform transmission measurements on tablets and translucent materials. The detector used for these measurements is (RT)-InGaAs;
e) fiber optic module with two probes. The detector used for FOP measurements is (TE)-InGaAs;
f) integrating sphere window for diffuse reflectance measurements. The detector used for IS measurements is (TE)-PbS;
g) spectral standard reference for FOP measurements.

In Bruker Optics MPA the integrating sphere module is designed for the direct measurements of sample having large area, by using a circular window with diameter of about 15 mm and it also allows the rotation of the sample in a manner to average the spectrum collected over all the sample surface. An internal computer controlled reference (a gold coated wheel that close the window during the background acquisition) makes automatic the background acquisition without the need to remove the sample. In Figure 2.4.2 the measurement of a solid sample by means of the IS sampling tool is represented; as can be seen in Figure 2.4.2 b), the detector is placed at 90° with respect to both sample and NIR source.

As for the fiber optic module, in MPA FT-NIR spectrophotometer, the NIR spectra acquisition is possible on liquids (by direct probe immersion at maximum 10 mm depth), powders and solid samples.

The fiber optic probe can be used both for reflectance and transmittance measurements (Pérez et al., 2013). In the reflectance mode, in the same cable some optic fibers transport the light beam towards the sample surface, while some other fibers detect the diffused light and carry it to the detector. On the contrary, in the transmittance mode, the optic probe is commonly used as flexible NIR radiation source that can be easily directed to a sample cell.

In this work, only diffuse reflectance measurements were performed.

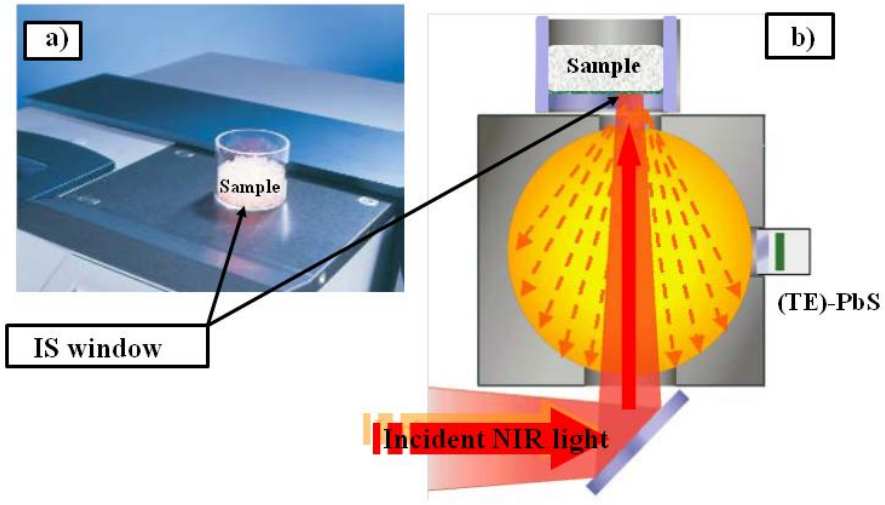


Figure 2.4.2. Measurement of a solid sample using the IS module (a) and detail of the reflection of the NIR radiation into the IS (b).

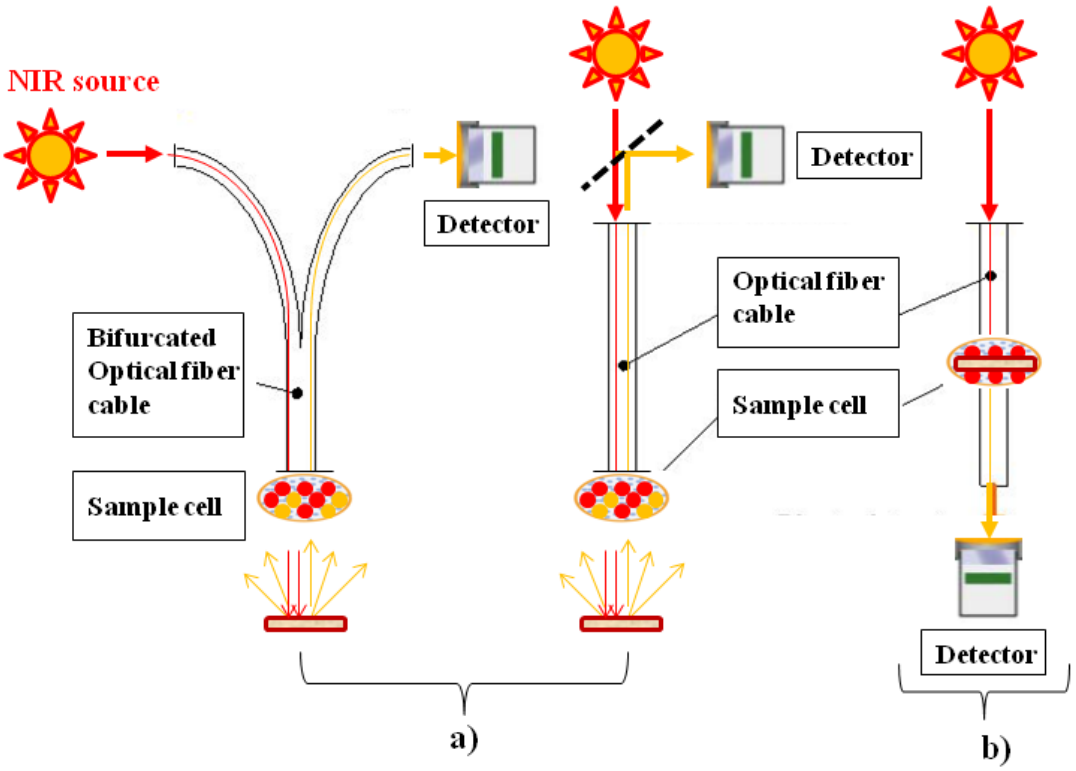


Figure 2.4.3. Fiber optic cables used for diffuse reflectance measurements (a) and for transmittance measurements (b).

2.5 Advantages and disadvantages of NIR infrared spectroscopy

As explained above, for the study of complex matrices, the use of NIR spectroscopy is recommended because, in particular, it allows to analyze unaltered and unchanged samples, and provides chemical information as simultaneous contributions about all the components and molecules present in the samples. NIR spectroscopy shows several advantages with respect to other chemical analysis and instrumental techniques used on complex matrices as food. First of all, it is a non-invasive and non-destructive technique: solid samples can be easily measured with no sample pretreatment, if an appropriate sampling tool is used. Indeed, the IS and FOP sampling tools used in this work required no sample preparation. As a consequence, in comparison to traditional analytical techniques, in NIR spectroscopy there is no need for reagents or particular materials/procedures to prepare the samples. In this way, the analytical costs result much lower. Then, the NIR analysis is simple to execute, so that the spectra acquisition can be recorded in very short time also by means of unexpert personnel. In the end, measurements and results deliveries are now rather fast: great developments in calculation power of computers, NIR instruments and chemometrics methods have enabled the real-time extraction of analytical responses from samples, allowing the quality control directly during the transformation processes in factory.

On the other hand, this technique also presents some disadvantages. One of the main problem is that, when NIR spectroscopy is used as fingerprint technique to gain calibration or classification models, it has to be considered as a relative method. In fact, accurate and robust models can be obtained only using a large number of samples in a manner to include all variations in physical and/or chemical properties. Moreover, the sensibility of NIR spectroscopy makes it generally applicable only to the determination of the major components of the sample.

2.6 References

Ben-Gera, I., Norris, K.H., (1968). Determination of moisture content in soybeans by direct spectrophotometry. *Israeli Journal Agr. Res.* 18, 124-132.

Berntsson, O., Danielsson, L.G., Folestad, S., (2001). Characterization of diffuse reflectance fiber probe sampling on moving solids using a Fourier transform near-infrared spectrometer. *Anal. Chim. Acta.* 431, 125-131.

Bokobza, L., (1998). Near Infrared Spectroscopy. *J. Near Infrared Spectrosc.* 6, 3-17.

Cozzi, R., Protti, P., Ruaro, T., (1997). Analisi Chimica Strumentale B, Ed. Zanichelli.

Coates, J., (2000). Interpretation of infrared spectra, a practical approach. Encyclopedia of Analytical Chemistry, Ed. Wiley & Sons Ltd.

Coates, J., Sanders, A., (2000). A universal sample handling system for FT-IR spectroscopy. *Spectrosc. Europe.* 12(5), 12-22.

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

MPA Bruker Optics FT-NIR spectrophotometer brochure, Bruker Optics Inc. Billerica, MA, USA
MPA Bruker Optics FT-NIR spectrophotometer user manual, 2nd updated edition (2007), publication date September 2007.

Olinger, J.M., Griffiths, P.R., Burger, T., in: Burns, D.A., Ciurczak, E.W., (Eds.). (2001). Theory of diffuse reflection in the NIR region. In: Handbook of Near-Infrared Analysis, Marcel Dekker, New York, US, pp. 19-52.

Perez, M.A., Gonzalez, O., Arias, J.R., Chapter 10 (2013): Optical fiber sensors for chemical and biological measurements. In: Current developments in optical fiber technology, Eds. Sulaiman, W.H. and Hamzah, A.

Rubinson, K.A., Rubinson, J.F., (2002). Chimica Analitica Strumentale, Ed. Zanichelli.

Workman, jr J.J., and Burns, D.A., (2001). Commercial NIR Instruments, In Burns, D.A., and Ciurczak, E.W., (Eds) Handbook of Near Infrared analysis (2nd Ed.), Marcel Dekker Inc, New York p. 53-70.

Chapter 3: CHEMOMETRIC METHODS

3.1 Multivariate analysis methods

What is chemometrics?

The characterization of a food product, considered as a complex matrix, requires considerable analysis/analytical measurements made on many samples. Normally these data sets are analyzed using the classical statistical approach that involves the analysis of only one or two variables at a time (defined as univariate or bivariate data analysis). However, this approach is not able to discover the relationships between all samples and all variables in an efficient manner; to do this, we have to simultaneously process all data.

Chemometrics is the science of extracting information from chemical data, using tools of statistics and mathematics, by considering all the features simultaneously, in a multivariate approach (Workman, 2002; Todeschini, 1998; Massart, et al., 1997). The relevant chemical information contained in a multivariate dataset (for example, a FT-NIR spectra dataset) can be took out by using multivariate data analysis methods which relate a number of analytical variables (such as the single wavelengths in the spectrum) to other chemical, physical or rheological variables of the sample (such as concentration). The multivariate data analysis can be used for different aims including the planning of the experiments, the data exploration to find particular patterns, the classification of the samples in different categories and the regression, used to determine specific properties of unknown samples.

Advantages and disadvantages of chemometrics

According to (Workman, 2002), the use of chemometric approaches applied on spectroscopic dataset presents clear advantages:

i. chemometrics increases the speed in obtaining real-time information from a dataset;

ii. it is able to extract high quality information from less resolved data (noisy spectra);

iii. it allows to provide resolution and discrimination power when applied to complex data;

iv. it supplies methodologies and diagnostics for the integrity and probability that the information it derives is accurate by using highly coherent approaches, logical, objective and

methodological techniques, in accordance with the scientific methods and experimental evidence.

v. it allows to improve rationally the data acquisition procedure by means of a proper design of experiments;

vi. it improves knowledge of existing dataset and processes;

vii. it has very low capital requirements.

In summary, chemometrics provides the promise of faster, cheaper and better information extracted from data. In addition, it is common sense to know that math is cheaper than physics. Indeed computer programs and new algorithms can solve problems that traditionally have required time and extensive hardware developments. Chemometrics represents a superior approach: science evolution demonstrates that intelligence can replace physical and material solutions as much as the digital chip replaces the mechanical instrument works.

The main disadvantages of chemometrics are closely related to the widespread lack of knowledge about what chemometrics is and what it can realistically perform. Indeed, even if the chemometrics studies are increased in the academic field, there are relatively few workers actually using it for solve their daily problems in industry. As a consequence, in the industrial context, there is a lack of official practices and methods associated with chemometrics.

In most of the cases, chemometrics is considered too complex for the middle technician. Even if mathematics involved in chemometric algorithms is not so simple, nowadays the chemometric softwares became really user-friendly and do not require high theoretical knowledge.

The other disadvantage is that chemometrics needs a change in the way of thinking. We live in a multivariate world but we are used to apply the scientific procedures in a univariate, or at least bivariate, way. It is not unlikely that a change of mentality may take more effort than getting comfortable with complex mathematical tools.

Chemometric approaches

The diffusion of powerful personal computers and the development of efficient software tools allow to process large amounts of data in a short time, making full use of information gathered in the lab. In fact, the modern instrumental analytical techniques, often automated, offer the possibility to rapidly collect huge amounts of data. In this current context, chemometric techniques based on multivariate data analysis have become necessary.

Chemometric methods for multivariate data analysis can be identified according to their aims and to the algorithms or procedures used. The selection of the optimum method depends on the purpose of the analysis, the characteristics of the samples and the complexity of the analytical data obtained (for example the non-linearity).

The chemometric approaches aimed at showing similarity/dissimilarity between samples are known collectively as "pattern recognition methods" which are defined as "supervised" or "unsupervised", according to the class to which the samples belong to is known or not known.

Principal Component Analysis (PCA) is one of the most widely used unsupervised methods. In the first instance, PCA overcomes the issue to graphically represent the original data without loss of useful analytical information. In fact, if original data, and the complex analytical information herein contained, can be defined by using at most three variables, a direct visual representation may be possible. Otherwise, when the number of variables is higher, the data display is not possible and the prospect to directly identify correlations among the analyzed samples becomes more difficult. PCA has been developed to contemporarily reduce the space dimensionality used to represent the data structure and to achieve an effective data visualization. This can be done by defining a different set of variables, most appropriate to identify patterns and to find the existing relationships among the objects, including the identification of some objects characterized by very strange behaviors compared to the others and defined as *outliers*.

Concerning the supervised methods, similarity is quantitatively verified on the basis of mathematical criteria and may be defined as the correlation coefficient between samples and/or making use of a distance measurement. The different kinds of classification methods available establish boundaries between the different classes, or they model the space occupied by a single class and determine whether a sample belongs to it on the basis of distance measurements or the residual variance.

Other multivariate methods have been developed to construct models capable of accurately predicting the characteristics and properties of unknown samples; these methods are named as multivariate calibration or regression methods, and they involve the steps described in Table 3.1.1.

Step	Description	Purpose
1)	Choosing the calibration samples	To select a set of samples representative of the whole population.
2)	Determining the target parameter by using the reference method	To determine the value of the measured property in an accurate, precise manner. The quality of the value dictates that of the calibration model.
3)	Recording the NIR spectra	To obtain physico-chemical information in a reproducible manner.
4)	Subjecting spectra to appropriate treatments	To reduce unwanted contributions (noise) to the spectra.
5)	Constructing the model	To establish the spectrum-property relationship using multivariate methods.
6)	Validating the model	To ensure that the model accurately predicts the property of interest in samples not subjected to the calibration process.
7)	Predicting unknown samples	To predict rapidly the property of interest in new, unknown samples.

Table 3.1.1. Steps in the multivariate model-construction process. As an example, calibration by means of NIR spectra has been used.

The multivariate regression method most frequently used in infrared spectroscopy is *Partial Least-Squares* (PLS) regression. PLS can be used considering both specific spectral regions or the whole spectrum, and it allows more information to be included in the calibration model. PLS finds the directions of greatest variability by considering both spectral and target-property information, with the new axes called "Latent Variables" or "PLS components" or "PLS factors". In some cases, however, the spectral data and the target property are not linearly related as a result of instrumental factors or the physico-chemical nature of the samples. These cases can be addressed using non-linear calibration methods or transforming the original variables by using a non-linear (logarithmic, exponential or quadratic) function and then doing the regression on the transformed variables. Once models are constructed, their predictive capacity must be checked on samples subjected to the same treatment (spectrum recording conditions and spectral pretreatment) as those used for calibration but not employed to construct the model.

3.2 Pretreatment of spectral data

The data pretreatment consists in the modification of original data (e.g., FT-NIR spectral dataset) before building a model or otherwise analyzing those data. The purposes of the pretreatment procedure is to obtain a linear response for the variables and to remove extraneous sources of variation, such as noise, which are not interesting for the aim of the works.

An original dataset of spectra can be considered as a matrix **X** {*i*, *k*}, formed by x_{ik} elements obtained by measuring *k* variables on *i* samples. In the specific case of a spectral dataset, each *i* spectrum is defined by the *k* intensities at all the recorded wavelengths.

Depending on the aim of the work, on the time available, and on the used software, advanced preprocessing procedures that include more than a simple pretreatment can be used. The pretreatment methods can be distinguished in *row-wise* and *column-wise*. The row-wise methods work on each sample at a time and are useful when it is necessary to remove unwanted variance from single samples. The column-wise methods, on the other hand, work on variables of the data, assuming that any variance in the data may be important. There are several types of pretreatment methods, a good choice among them being only possible by critical observation of the data as a whole, but generally it is recommended to perform row-wise pretreatments before to any column-wise. Finally, it has to remember that each single pretreatment has advantages and disadvantages since it is usually not possible to establish in advance the one working better (Rinnan et al., 2009). In any case, the pretreatment choice is a compromise in which different parameters are involved. For instance, some row-wise pretreatments can modify the position of interesting peaks and so interfere with the spectral interpretation (Fearn et al., 2009).

Meancentering and autoscaling

Starting from column-wise ones, the most common and simple pretreatment procedures are the scaling methods represented by meancentering and autoscaling.

In meancentering, the values of each variable are subtracted by the mean. The use of variable meancentering makes the model results unaffected by the variable average values.

As a step forward, autoscaling is appropriate when large differences in the variance of the different variables are present, especially if the variables in the system are expressed by different units, which are supposed to be of comparable meaning. Autoscaling consists in centring the data by subtracting the mean value and dividing by the standard deviation.

In a multivariate sense, the scaling procedures translate the collection of data to the origin of the multivariate space where analysis will be performed. The practical consequence of the scaling procedure is often a more simple and interpretable calibration and classification model.

The meancentering procedure provides data that are centered with respect to each variable mean value

$$x'_{ik} = x_{ik} - x_{m,k} \qquad (3.1)$$

The fundamental property of the meancentered data is that the mean value of each variable is zero: $x'_{m,k} = 0$ (3.2)

This kind of scaling does not modify the variance of the data.

The autoscaling procedure include an additional step with respect to meancentering, in fact it consists in the centering of the data followed by a normalization to unit variance:

$$x'_{ik} = (x_{ik} - x_{m,k})/S_k \qquad (3.3)$$

Mean value and variance of the autoscaled variables are respectively 0 and 1:

$$x_{m,k} = 0 \quad \text{and} \quad V(x'_k) = 1 \qquad (3.4, 3.5)$$

The term $V(x)$ indicates the variance and s_j is the standard deviation of the j variable.

For datasets including variables expressed in different unit scale autoscaling is recommended, while for highly correlated variables (i.e., spectroscopic absorptions at different wavelength) the pretreatment of FT-NIR spectra by meancentering is strongly suggested.

The different effect produced on a set of spectra by meancentering and autoscaling is shown in Figure 3.2.1.

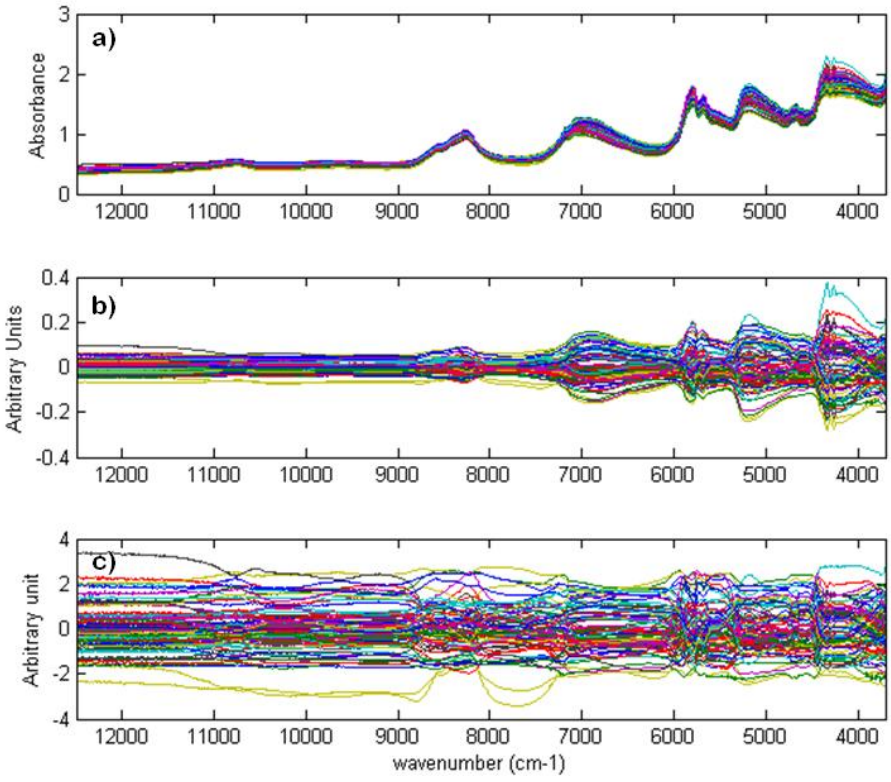


Figure 3.2.1. FT-NIR spectra of pig fat samples: a) original spectra, b) meancentered spectra, c) autoscaled spectra.

Smoothing

Smoothing is an effective procedure to reduce the spectra noise, when the averaging of multiple measurements is not possible or sufficient. The drawback is that a sub-optimal cut-off frequency will also influence the useful high-frequency signals (narrow peak in FT-NIR spectroscopy) by reducing the spectral resolution. These effect have to be balanced by a proper algorithm that use a proper mobile window as filter signal. The most simple filtered signal is obtained by a selected width window that calculate an arithmetic mean of *n* discrete signal values within the corresponding spectral window and then move to the next variable until all the NIR spectral wavelength is covered. This approach was proposed by Savitzky and Golay in 1964 (Savitzky and Golay, 1964) and now represents a common algorithm for the smoothing pretreatment. The algorithm essentially fits individual polynomials to windows around each point in the spectrum. The mathematical functions used to smooth the data are polynomial functions. The effect of smoothing by interpolation of data is carried out by means of moving window. In fact, the algorithm requires both the selection of the window size (width of the filter) and the order of the polynomial function (linear, quadratic, cubic, etc.). Typically, the window width should be of the order of, or less than, the nominal width of non-noise features. In the PLS Toolbox for MATLAB© platform, the default setting for the Savitzky-Golay smoothing are adjusted to a window size (filter width) of 15 points, using a polynomial function order equal to zero and derivative order equal to zero.

First and second derivatives

Derivatives of spectra are used to remove or suppress constant background signals and to enhance the visual resolution. Background signals and global baseline variations are low-frequency phenomena, so derivatives can be interpreted as high-pass filters.

There are several ways to perform the derivative of a spectrum: the simplest way is to perform a point-difference first derivative, in which each point of the sample is subtracted from the point immediately neighboring. In this way, the signal that is the same between the two variables is removed, while are left only the different parts of the signal. The result of the first derivative point-difference applied to the entire sample is the removal of any offset from the sample and the decrease of the “importance” of the lower frequency of the spectrum. The second derivative is calculated by doing again the procedure, so emphasizing the higher frequency features.

Compared to the original spectrum, the first derivative spectrum is the slope at each point of the original spectrum. It has peaks where the original has maximum slope, and crosses zero at peaks in the original. The second derivative is the slope of the first derivative. In some way the second derivative spectrum is more similar to the original one, having peaks in roughly the same places, although they are inverted in direction. As visual effect, the second derivative transfers peak maxima into minima and vice versa. It is a measure of the curvature in the original spectrum at each point. The fact that the measured spectrum is not a continuous curve, but a series of measurements at equally–spaced discrete points is taken into account. Similarly to the smoothing technique, the Savitzky-Golay algorithm can be used for the calculation of derivative as well, following the mobile window procedure. Also for the derivative process the algorithm requires arbitrary selection of the size of the window (filter width), the order of the polynomial, and the order of the derivative. The choice of the mobile window size involves a trade-off between noise reduction or smoothing (for which a wide window is preferred in high resolution spectra) and distortion of the curve, which could happen if the window is too wide. The polynomial functions used to fit the spectra in the mobile window could be selected evaluating noise reduction and spectral distortion; however, since this step could be time-consuming, using the default setting of chemometric software is an attractive option in these circumstances.

In the PLS Toolbox for MATLAB© platform the default setting for the Savitzky-Golay derivative are adjusted to a window size of 15 points, using a second order polynomial function and derivative order equal to one for first derivative and 2 for second derivative. An important aspect of derivatives is that they are linear operators, and indeed, do not affect any linear relationships within the data.

Figure 3.2.2 shows the results of first and second derivatization on NIR spectra of pig fat samples using the default Savitzky-Golay derivative setting. Notice that the first derivative operation has removed the predominant background variation from the original data. The smaller variations remaining are due to the chemical differences between samples. Side effects of derivatives on spectroscopic data are the loss of the original shape of the spectral curve and the reduction of the signal-to-noise ratio.

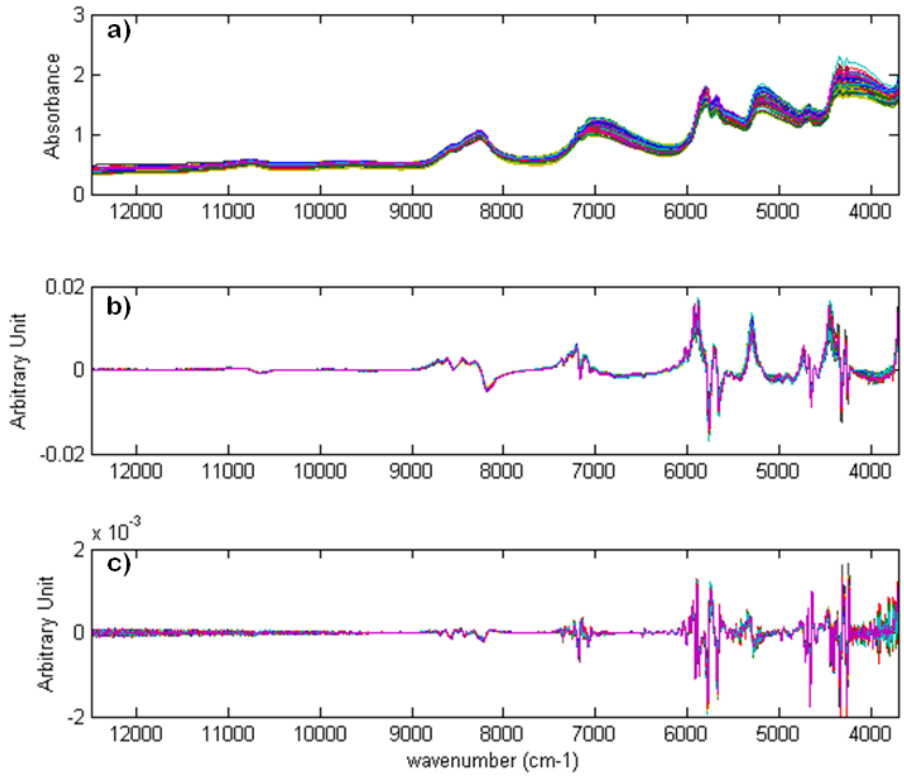


Figure 3.2.2. FT-NIR spectra of pig fat samples: a) original spectra, b) spectra after first derivative, c) spectra after second derivative.

Detrend

Sometimes a collection of spectroscopic data may denote the presence of a constant, linear or curved offset that can be eliminated or minimized through the use of the signal pretreatment called detrend (Barnes et al., 1989).

The detrend pretreatment procedure consists to fit a low-degree polynomial through all data points of the original spectrum and then subtract the resulting function's curve from the spectrum. Although the use of spectral pre-treatments are often necessary and useful, some of them change and confuse even part of the useful information present in the spectra. Detrend minimizes the useless variation for modeling, but in some cases could produce non-linear responses from linear ones. Furthermore, the fact that a single polynomial is suitable for each signal can augment the interfering variance in the dataset. For these reasons, detrend has to be used only when the whole signal and background are highly similar to one another and when the signal is not strongly influenced by physico-chemical processes under investigation.

Figure 3.2.3 shows the effect of detrend on NIR data acquired on pig fat samples. Notice that, using first order detrend, the general linear offset is removed from all spectra, giving a mean intensity of zero for each spectrum.

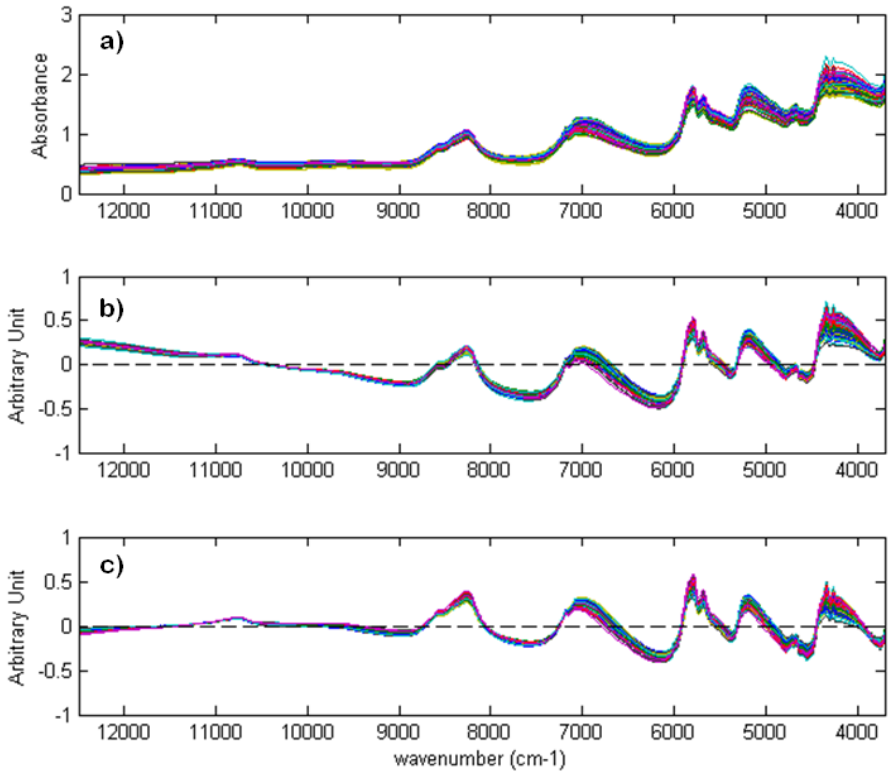


Figure 3.2.3. FT-NIR spectra of pig fat samples: a) original spectra, b) spectra after first order detrend, c) spectra after second order detrend.

Standard Normal Variate and Multiplicative Scatter Correction

The spectrum of a solid sample is influenced by the physical properties of the matrix. In fact light scattering produced from solid samples (powder), as well as emulsion and dispersions result in multiplicative deviations. Other frequent sources of multiplicative perturbations may be changes in the optical pathlength and sometimes in the sensitivity of the detector and amplifier in the analytical spectrophotometer. These effects are difficult to approximate with linear factor combinations during the calibration process.

This poses some problems in evaluating properties of samples which are not dependent on the physical appearance (such as analysis of raw materials and determination of composition). In these situations, spectral pretreatment are needed to minimize those contributions incorporating irrelevant information and in order to be able to develop more

simple and robust models. Among these pretreatments, Standard Normal Variate (SNV) (Barnes et al., 1989; Guo et al., 1999) and Multiplicative Scatter Correction (MSC) should be mentioned (Martens et al., 1989; Geladi et al., 1985; Isaksson and Naes, 1988).

More in detail, the SNV method aims to remove the multiplicative effect of scatter and particle size on an individual object basis. The transformation is carried out according to the following formula:

$$\hat{x}_i = \frac{(x_i - \bar{x})}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}} \qquad (3.6)$$

where x_i represents the percent transmittance value (corresponding to the point *i* of the spectrum), $\bar{x}$ is the mean value calculated over the percent transmittance of all the spectral points, *n* is the number of the points that constitute the spectrum and $\hat{x}_i$ is the value that assumes the intensity of the spectrum in the *i* point after the application of the SNV transform. The actual filtering can be referred to as meancentering and scaling to unit variance in the object direction. The SNV is insensitive to any deviating spectra, since the correction is independent of average spectrum.

Otherwise, the MSC method is a processing step that attempts to account for scaling effects and offset (baseline) effects (Martens et al., 1989). This correction is achieved by regressing a measured spectrum against a reference spectrum and then correcting the measured spectrum using the slope (and possibly intercept) of this fit.

Specifically, MSC follows this procedure. Define $\boldsymbol{x}$ as a column vector corresponding to a spectrum to be standardized and $\boldsymbol{r}$ as a vector corresponding to the reference spectrum (often this is the mean spectrum of the calibration data set). The vectors are most often mean-centered according to:

$$\boldsymbol{x}_c = \boldsymbol{x} - x_m \, \mathbf{1} \qquad (3.7)$$

$$\boldsymbol{r}_c = \boldsymbol{r} - r_m \, \mathbf{1} \qquad (3.8)$$

where $\boldsymbol{x}_c$ and $\boldsymbol{r}_c$ are the mean centered vectors, x_m and r_m are the respective means, and $\mathbf{1}$ is a vector of ones. The unknown multiplicative factor b is determined using:

$$\boldsymbol{x}_c = b \, \boldsymbol{r}_c \qquad (3.9)$$

$$b = (\boldsymbol{r}_c^T \, \boldsymbol{r}_c)^{-1} \boldsymbol{r}_c^T \, \boldsymbol{x}_c \qquad (3.10)$$

and the corrected spectrum $\hat{x}_i$ is then given by:

$$\hat{x}_i = \frac{\boldsymbol{x}_c}{b + r_m \mathbf{1}} \qquad (3.11)$$

The effects of SNV and MSC pretreatment procedures are shown in Figure 3.2.4, in comparison with the original un-pretreated dataset.

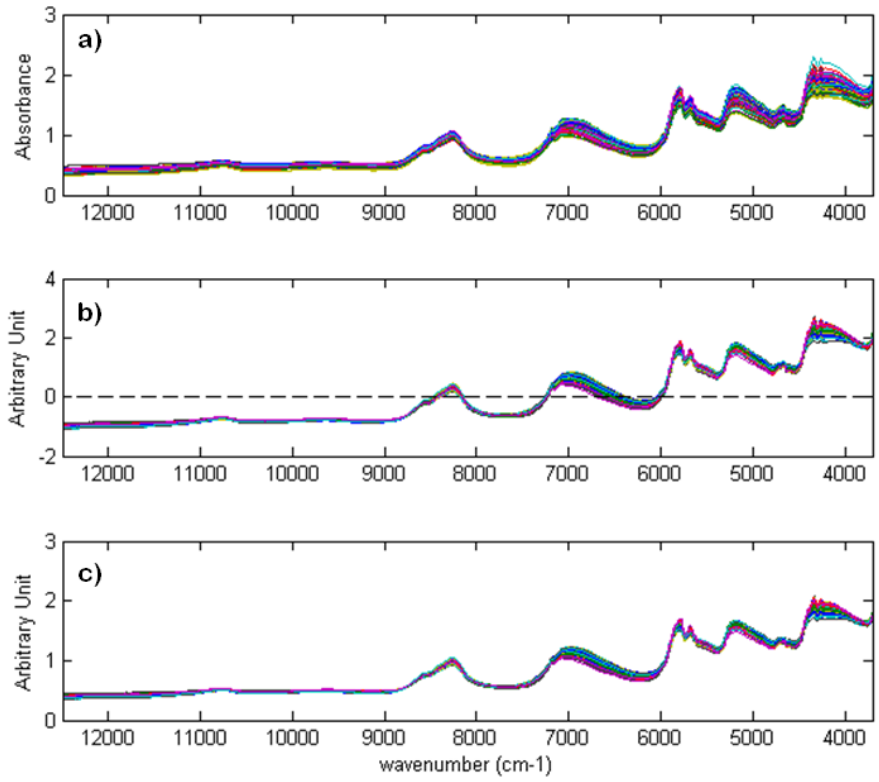


Figure 3.2.4. FT-NIR spectra of pig fat samples: a) original spectra, b) spectra after SNV, c) spectra after MSC.

3.3 Principal Component Analysis

Principal Component Analysis (PCA) is an efficient unsupervised pattern recognition technique. Its goal is to identify 'structures' or 'patterns' (correlations) among the objects constituting the system under study, without making any *a priori* assumptions. Hence, it constitutes a tool for identifying patterns in the objects ensemble and for picturing them, in order to evidence at best both similarities and differences. It also gives a way to 'compress' the representation of the objects in the original k-dimensional data structure (with k = number of original experimental variables) into a smaller A-dimensional space (with A = number of 'new' selected variables, i.e., Principal Components, in brief PCs), without loss of useful information but, rather, by visualising it much better.

To this purpose, PCA computes suitable linear combinations of variables, that describe the major trends in the ensemble of the data. Essentially, a set of partially correlated variables is transformed into a new set of completely uncorrelated variables, the PCs, which are ordered

according to the progressively lower variability they account for. These PCs lack of any physical meanings, being linear combinations of the original, physically meaningful variables. Most often, negligible useful information is retained from 4th or 5th principal component variable on, so that the system can be profitably represented in a 2- or 3-dimension space. The original variables give different contributions to the different principal components, corresponding to the importance they have in defining the direction of the PCs. For each PC, the contribution of the original variables is contained into the corresponding *loadings* vector, essentially consisting of the coefficients of the original variables in the linear combination: the higher is the absolute value of a given coefficient, the higher the importance of the corresponding original variable for that PC. At the same time, the position of each object on a given PC is given by the corresponding *scores* vector.

The aim of the PCA is to divide the total variance of the data matrix **X** in a first part associated to the variables *k* (*loadings*, p_{ak}), in a second part associated to the objects *i* (*scores*, t_{ia}) and in a third part that describes the non-systematic variation (*residuals*, e_{ik}):

$$x_{ik} = x_{m,k} + \Sigma(t_{ia}\, p_{ak}) + e_{ik} \qquad (3.12)$$

where the sum is extended to the number of the extracted principal components. The mean value of each variable ($x_{m,k}$) is the reference point for the model variation.

The decomposition may be schematically represented as in Figure 3.3.1.

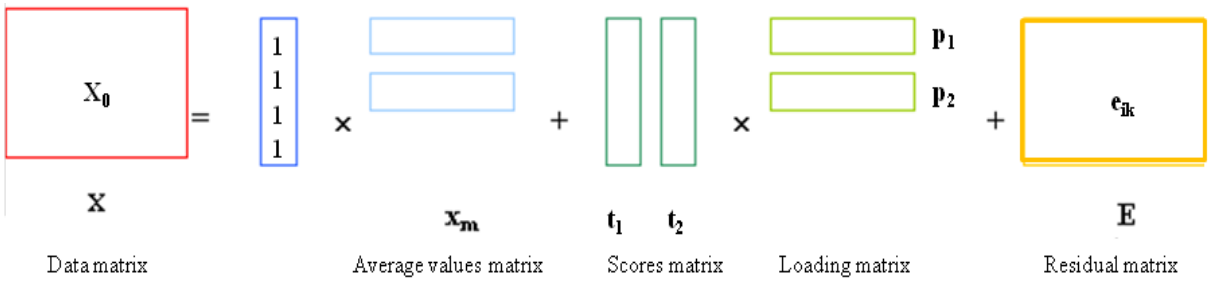


Figure 3.3.1. Data matrix decomposition carried out by PCA.

In particular, the PCA is a multivariate statistical method which considers that the variables that define an object characterize it simultaneously, with reciprocal influences and correlations. The basic assumption of this model is that the objects, on which the same variables are measured, can be essentially considered similar, hence it is right to create an approximate similarity model, only locally valid.

The main goals of the PCA are:

i. To evaluate the real dimensionality of the data, i.e. the number of latent variables (or the number of significant principal components, A);

ii. To provide a graphical representation of the relationships among objects (score plot) and variables (loading plot) in the data matrix, so that the systematic trends, the deviations, the clusters and the correlations are emphasized and the objects with anomalous behavior (outliers) are identified.

From a geometrical point of view, PCA is a method that projects the data from a original variables space in a reduced space, defined by A-PC dimensions. Since both the loadings and the scores vectors are orthogonal, they can be used as coordinate axes. PCA, in fact, searches for directions of maximum variability of samples in variables space and uses them as new orthogonal and rotated axes (Wold et al., 1987; Jackson et al., 1991). Considering, for example, the two-dimensional case (Figure 3.3.2), the first new dimension falls on the longest axis of variance; while the second new dimension contains the rest. As much variance as possible is forced into x-prime. The rest is spanned by y-prime. Mathematically, this is a simple linear algebra transformation for any reasonable number of dimensions.

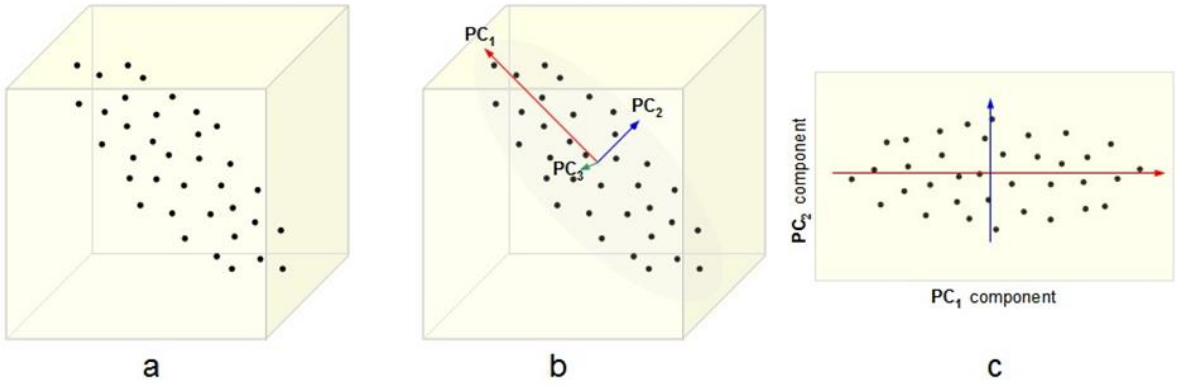


Figure 3.3.2. Individuation of the PC space with respect to the original domain: objects in the original domain (a), individuation of the PC space (b), objects projected in the PC space (c).

A set of statistical parameters is calculated within the PCA model; among them the *Residuals Q* and *Hotelling* T^2 are of great importance.

Q is simply the sum of squares of each row (sample) of the residual matrix **e**, *i.e.* for the i^{th} sample in **X**, $\mathbf{x}_i$:

$$Q_i = \mathbf{e}_i\mathbf{e}_i^T = \mathbf{x}_i(\mathbf{I} - \mathbf{P}_k\mathbf{P}_k^T)\mathbf{x}_i^T \qquad (3.13)$$

where $\mathbf{e}_i$ is the i^{th} row of $\mathbf{E}$, $\mathbf{P}_k$ is the matrix of the k loadings vectors retained in the PCA model (where each vector is a column of $\mathbf{P}_k$) and $\mathbf{I}$ is the identity matrix of appropriate size (n by n). The Q statistic indicates how well each sample conforms to the PCA model. It is a measure of the difference, or residual, between a sample and its projection into the k principal components retained in the model.

The sum of normalised squared scores, i.e., the Hotelling T^2 statistic, is defined as:

$$T_i^2 = \mathbf{t}_i\lambda^{-1}\mathbf{t}_i^T = \mathbf{x}_i\mathbf{P}\lambda^{-1}\mathbf{P}^T\mathbf{x}_i^T \qquad (3.14)$$

where $\mathbf{t}_i$ in this instance refers to the i^{th} row of $\mathbf{T}_k$, the matrix of k scores vectors from the PCA model and λ^{-1} is the diagonal matrix containing the inverse of the eigenvalues associated with the k eigenvectors (principal components) retained in the model.

In the context of a geometric interpretation, Q is a measure of the variation of the data outside of the principal components plane defined in the PCA model. As it can be seen in Figure 3.3.3, Q is a measure of the distance out of the plane formed by the 2 PC model of the data point. A sample with a large Q is shown on the upper left side of the figure. This sample is out of the plane of the model (although its projection into the model is not unusual). The Q limit defines a distance out of the plane that is considered unusual based on the data used to form the PCA model.

T^2, on the other hand, is a measure of the distance from the multivariate mean, i.e., the intersection of the PCs in the figure, to the projection of the sample onto the 2 PCs. The T^2 limit defines an ellipse on the plane within which the data normally project.

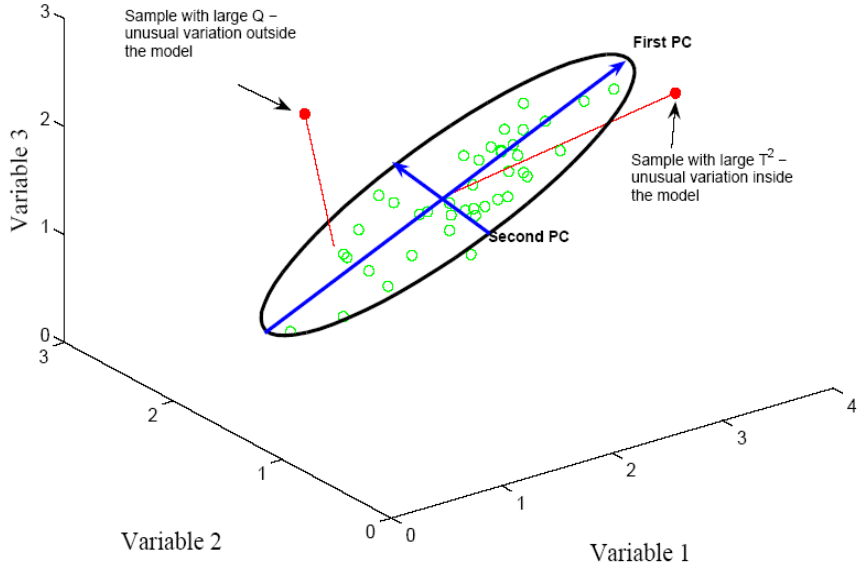


Figure 3.3.3. Individuation of the model space around the samples.

It is important to note that the confidence limits defined for T^2 and Q are based on the assumptions that the data distribution are multivariate normal. Obviously, experimental measurements are often not normally distributed, but the assumption is more evident and real for random errors (represented in residual Q parameter). Moreover, regarding high dimensions dataset (e.g., spectroscopic dataset, including hundreds of samples and thousand of variables), the data normal distribution is still and more valid. Indeed, the central limit theorem states that sums of several different groups will tend to be normally distributed, regardless of the probability distribution of the individual groups (Anderson, 1984). The theorem suggests that reduced space methods such as PCA will tend to produce measures that are more normally-distributed than the original data.

The Q-T^2 plots allow the direct comparison of samples, in other words, reveal if a sample is 'similar' to all the other samples and belongs or not to the same normal distribution population. In these plots is possible to identify for which of the two parameters the samples are (or are not) within the 95% of confidence limit to belong to the same population. To make a comparison with the classical statistics, the limit of 95% corresponds to the portion of the Gaussian distribution defined by the mean values ± 2 standard deviations, while the confidence limit of 99.7% corresponds to the portion of the Gaussian distribution defined by the mean values ± 3 standard deviations. So, the confidence limit can be used as a sort of threshold for outlier determination: each sample outside the confidence limit is considered anomalous.

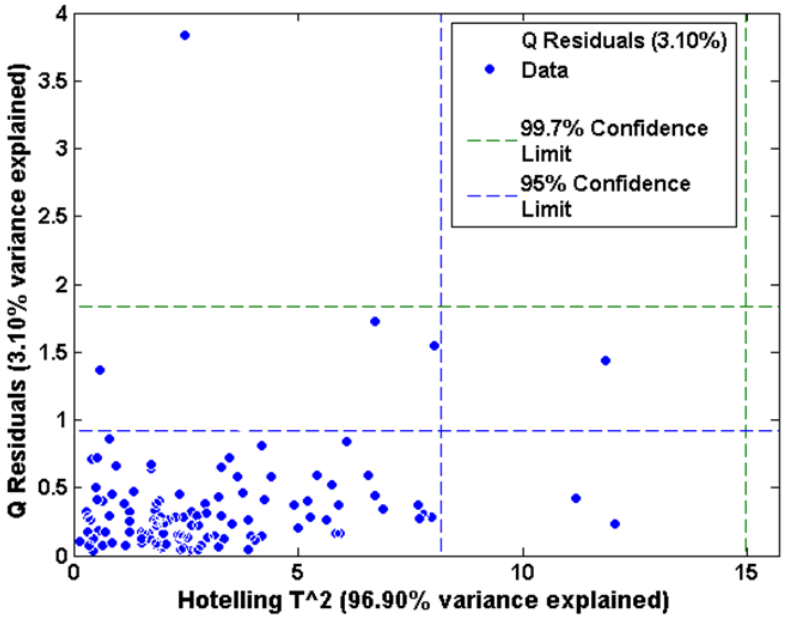


Figure 3.3.4. Example of a Q-T^2 plot obtained on a FT-NIR spectra dataset.

For instance, in Figure 3.3.4 the Q-T^2 plot obtained on a mean spectral dataset acquired by IS on pig fat samples is reported. In this figure it is possible to evaluate the presence of some outliers: none sample present anomalous T^2 values (higher than the 99.7% confidence limit), while three samples presents T^2 values greater than the 95% confidence limit. These samples should be considered anomalous for their position inside the model and they may be candidates for outlier selection. Concerning the Q residuals parameter, with the exception of one sample, all the other samples are included inside the 99.7% confidence limit, while five samples are outside of the 95% confidence limit. Also these samples, presenting anomalous positions outside the model, are candidates for outlier elimination. The proper procedure to delete outlier samples consists in the progressive elimination of the samples showing the highest T^2 values, outside the desired confidence limits, and then of the samples showing the highest Q values. After the deletion of outlier samples the model recalculation is required.

3.4 Partial Least Squares regression

The essence of the projection methods, such as PCA and PLS, is to use the latent structure in the data in order to find the true relationships within and among blocks. In PLS (Martens et al., 1989; Geladi and Kowalski, 1986; Geladi, 2002; Wold et al., 2001; Martens, 2001) the object variation in the predictor block is described by the X-scores, **T**, and the corresponding variation in the response block is described by the Y-scores, **U**.

While PCA identifies the PCs as those directions describing the maximum variance of the X block, similarly PLS searches for a new set of variables, named Latent Variables (LVs), which maximise the covariance between the **X** and **Y** matrices, i.e., which both capture variance and achieve correlation. Definitively, PLS extracts those latent variables which separate from noise the information in the descriptor matrix **X** useful to predict the correlated part of the response matrix **Y**.

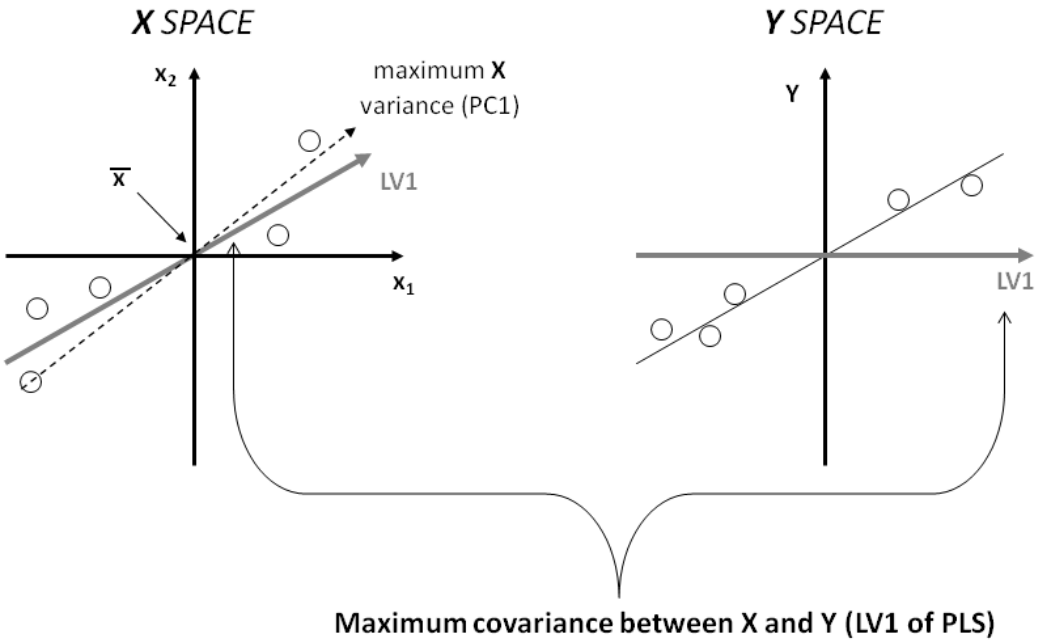


Figure 3.4.1. Relationship between **X** and **Y** in a PLS model.

For each dimension, a weight vector, **w**, is calculated, which contains the contribution of each **X**-variable to the explanation of **Y**, in that particular dimension. The matrix of weights, **W**, is crucial since it contains the structure in **X** that maximizes the covariance between **T** and **U** in each dimension. The matrices of X-loadings, **P**, and of Y-loadings, **Q**, are calculated for each dimension in order to perform the appropriate decomposition of **X** and **Y**. Hence, the decomposition of **X** and **Y** can be described as:

$$\boldsymbol{X} = \boldsymbol{T}\,\boldsymbol{P}' + \boldsymbol{E} \qquad (3.15)$$

$$\boldsymbol{Y} = \boldsymbol{U}\,\boldsymbol{Q}' + \boldsymbol{F} \qquad (3.16)$$

and the inner relation hold:

$$\boldsymbol{U} = \boldsymbol{T}\,\boldsymbol{b} \qquad (3.17)$$

Furthermore, the relations **T** = **XW** and **W** = **XU** are used iteratively in order to reach convergence.

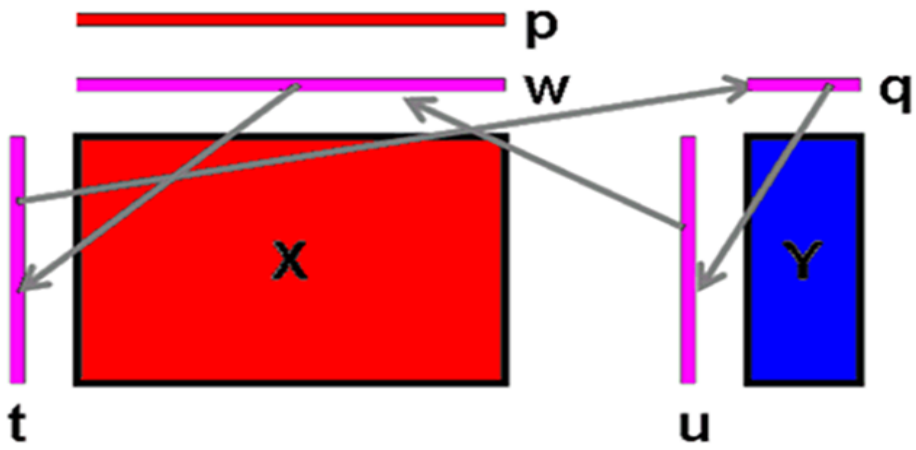


Figure 3.4.2. Decomposition of ***X*** and ***Y*** matrices computed by PLS regression algorithm.

The PLS regression coefficients can be computed using the equation:

$$\boldsymbol{B = W\,(P'\,W)^{-1}\,Q'} \qquad (3.18)$$

where the **W** is the matrix of the loading weight.

The estimate of **Y**, $\hat{\mathbf{Y}}$ is then given by:

$$\boldsymbol{\hat{Y} = X\,W\,(P'\,W)^{-1}\,Q' = X\,B + F} \qquad (3.19)$$

In which the error matrix **F** related to the prediction of response **Y** has to be minimized as much as possible, in order to realize a linear PLS prediction.

PLS1 and PLS2

The PLS regression algorithm is able to calculate calibration models both on response y as a single vector or on a **Y** matrix including different experimental responses. The PLS regression procedure applied on a single response vector y is called PLS1, while the same method applied on a response matrix **Y** is called PLS2; both the methods bring advantages and disadvantages.

Generally PLS1 leads to more precise results, especially when the y variables of the **Y** block are independent one another (i.e., they are orthogonal): this implies the calculation of more models, but the choice of the dimensionality of each single model may vary, making the procedure more flexible; also the interpretation of the results is generally easier. Conversely, PLS2 procedure is particularly faster when several y variables have to be predicted, since one

has to calculate only one single model. Moreover, when the y variables are correlated, one may benefit from the intercorrelations between them, diminishing the effect of experimental errors. Unfortunately, it is often difficult to obtain a single model which contemporarily performs well on a set of different responses.

Validation methods

In order to evaluate the predictive capability of a calibration model, two principal categories of methods can be employed: the internal methods and the external methods.

Among all the internal validation methods, that use the samples already present in the calibration model, the most common technique is *cross-validation* (Stone, 1974), in which the data set is divided in a certain number of groups (named cancellation or deletion groups). The calibration models are computed with the exclusion of one group at a time, and the excluded group is then predicted. The procedure is iterated since every group has been excluded once.

The different cross-validation procedures are characterized by different ways to select the deletion groups. A brief summary of the most common cross-validation procedures follows:

- Leave One Out (LOO), that is probably the most common and widespread cross-validation method; in LOO one single object is deleted at a time, then it is predicted. More severe techniques consider the deletion of groups of objects rather than only one.
- Contiguous Blocks cross-validation (CB) considers deletion of contiguous groups of objects. When samples have been ordered following a given criterion (e. g, sampling time, or increasing value of Y), it may be advantageous to choose the deletion groups in a different manner.
- Venetian Blinds cross-validation (VB): the objects are deleted in a way that each group "spans" all the rows interval.
- Random Groups (RG): though being more computationally intensive, is the more severe approach, which better mimes the performance in prediction of really unknown objects. RG cross-validation technique works on more cross-validation iterations.
- Custom cross-validation (Custom CV): a defined cross-validation vector is built by using selected groups of object directly chosen by the user.

The total sum of the square values of the differences between the predicted and the observed values is indicated as PRESS (Predicted Residual Error Sum of Squares), that is an index of the predictive capability of the model for a certain data set:

$$PRESS = \sum(y_{PRED} - y)^2 \qquad (3.20)$$

If the PRESS value is converted in an adimensional term, the cross-validated correlation coefficient is obtained. In analogy with the squared correlation coefficient, it is indicated as R^2_{CV}, i.e., the complementary value with respect to the ratio between the unexplained variance and the total variance:

$$R^2_{CV} = 1 - \frac{PRESS}{SS_Y} \qquad (3.21)$$

where SS_Y is the sum of squares of the original *Y* block.

R^2_{CV} represents an optimistic estimation of the prediction model ability and assess its robustness.

The error in prediction for each sample of the calibration set, i.e., the samples used to build the model, after the cross-validation procedure may be also obtained, and its value is expressed as Root Mean Squared Error of in Cross Validation (RMSECV), that is defined as follows:

$$RMSECV = \sqrt{\frac{\sum (y_{PRED-CV} - y)^2}{N}} = \sqrt{\frac{PRESS_{CV}}{N}} \qquad (3.22)$$

where N represents the number of samples of the calibration set (also called training set). The RMSECV values express the error in the same units of *y*.

Cross-validation is generally employed to define the correct model dimensionality and "complexity", i.e., determining the exact number of significant component (Latent Variables, LVs) that should be used to model the original variables. Selecting the optimal model dimensionality consists in checking whether the model has accounted for all the possible useful information and if noise has been discarded.

While the calibration error (i.e., the error in calculating the objects included in the model) ever diminishes with increasing the number of LVs, the error in cross-validation (RMSECV) reaches a minimum in correspondence with the optimal dimensionality, then it increases.

The actual predictive capability of a model is nevertheless verified by means of the external validation, by using a set of samples with known y-values (named test set), not used for the building of the model itself. The error in prediction for each sample of the external set is so obtained, and its value is expressed as Root Mean Squared Error of Prediction (RMSEP), that is defined as follows:

$$RMSEP = \sqrt{\frac{\sum (y_{PRED-EXT} - y)^2}{N}} = \sqrt{\frac{PRESS_{EXT}}{N}} \qquad (3.23)$$

where N represents the number of samples of the test set. Also the RMSEP value expresses the error in the same units of y.

As well as for the internal validation, also for the external validation is possible to estimate the prediction ability of the model, in this case considering as validation set the unknown objects included in the test set; using equations analogous to (3.20) and (3.21), the correlation coefficient in prediction R^2_{PRED} is calculated.

3.5 Partial Least Squares-Discriminant Analysis

While PCA describes the behaviour of a single data matrix **X**, the Partial Least Squares method (PLS) is devoted to the study of the relations between a matrix of *independent* variables **X**, and a vector/matrix of *dependent* (also called *predictor* or *response*) variables **Y**. As explained above, when there is only a single response variable, i.e., **y** is a vector, this method is named PLS1, while if **Y** is a matrix, the method is named PLS2 (1 and 2 being the dimensions of the response block in the two cases).

In the case of PLS1, by a logical point of view, the algorithmic procedure can be divided in two steps: firstly, PCA is performed onto the **X** matrix, and then the extracted PCs are put in relation to the **y** vector by the least squares method. In the PLS method, these two operations are executed contemporarily. In the case of PLS2, PCA is applied also to the **Y** matrix.

In other words, the selection of a few latent variables (operated by PLS algorithm) permits to identify only the useful information in the **X** block able to predict the correlated part of the response matrix **Y**. In extreme synthesis, using the equation (3.19), a linear calibration model is obtained. Once properly validated (by means of internal cross-validation and/or of an external test set), this regression model can be used to predict the values of the **Y** variable(s), given the corresponding values of the **X** descriptor variables.

PLS-Discriminant Analysis (PLS-DA) consists in a classical PLS2 where the **Y** response variable is a categorical one (replaced by the set of dummy variables describing the categories/classes) expressing the class membership of the statistical units (Wise et al., 2007). Therefore, PLS-DA algorithm develops a classification model that predicts the class number for each sample (Barker and Rayens, 2003).

To this aim, a categorical **Y** matrix is built, which contains as many variables (columns) as the number of modelled classes: the first column is the vector for class 1, the second column is the vector for class 2, and so on, as represented in Figure 3.5.1. For each column, the value of each variable is 1 if the sample belongs to a class, and 0 otherwise. For example,

in the case of a problem involving 4 classes, if the object 1 in the first row belongs to class 1, the corresponding first row in the **Y** matrix will be the following one: i.e., [1 0 0 0], only the element in the 1st column will be non-zero.

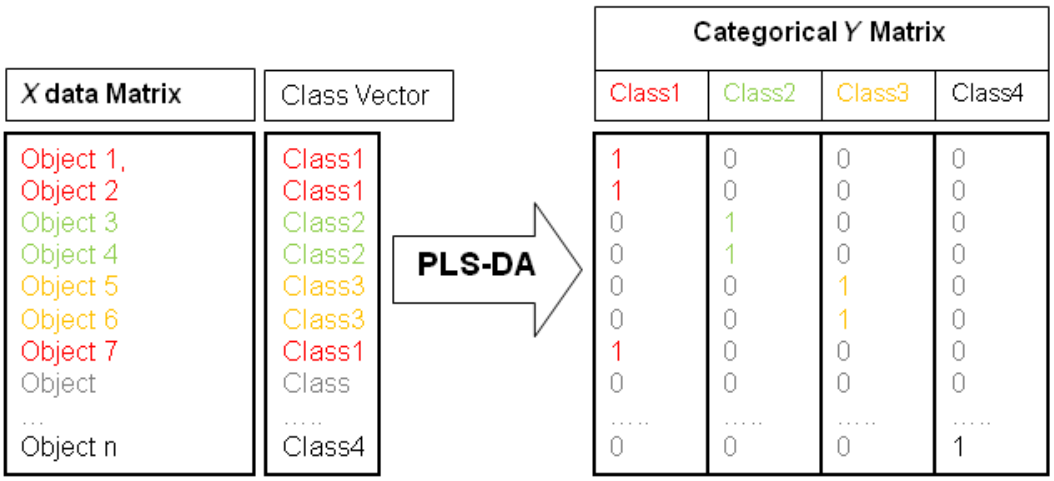


Figure 3.5.1. Categorical **Y** matrix built in a PLS-DA classification model.

Of course, the class values predicted by the PLS-DA model will not be exactly equal to 0 or 1. Therefore, a threshold value must be kept for each class model (i.e., for the prediction of each single class) so that an object having value greater than the threshold value is retained by the class model, while if its value is lower than the threshold it is rejected.

In order to define a proper threshold value, the Bayesian threshold calculation assumes that the predicted y values follow a distribution similar to what will be observed for future samples. Using these estimated distributions, for each class j a threshold is selected at the point where the two estimated distributions (i.e., the distribution of the 'class j objects' and the distribution of the 'not of class j objects') cross. This is the y-value at which the number of false positives and false negatives should be minimized for future predictions (assuming that the distribution of predicted y-values for a class is approximately normal).

Once built the model, it can be validated estimating its performance in cross-validation and in prediction of an external test set

The performance can be expressed in terms of Sensitivity (SENS, i.e., the percentage of objects of each class accepted by the class model) and Specificity (SPEC, i.e., the percentage of objects of the other classes rejected by the class model). According to (Forina et al., 2009) also the percentage Efficiency (EFF) can be calculated as the geometric meaning of SENS and SPEC for a single class. For a two class problem:

$$EFF = \sqrt{(SENS * SPEC)} \qquad (3.24)$$

3.6 Features selection

Spectroscopic signals are usually vitiated by an high degree of correlation of the variables. As a consequence, both in calibration and in classification ambit, notwithstanding the use of compression methods (such as the methods based on the Latent Variables) it is often convenient to operate a variables selection. As a result, a number of multivariate-analysis methods rely on feature reduction techniques that allow the dimensions of the original data to be reduced to a few uncorrelated variables containing only relevant information about the samples (Andersen and Bro, 2010; Xiaobo et al., 2010). During the feature selection step, particular attention must be paid in order to deal with a minimum number of variables without any loss in the predictive ability. A brief summary of the features selection methods includes interactive variable selection (Lindgren et al., 1994), uninformative variable elimination (Centner et al., 1996), iterative predicting weighing PLS (Forina et al., 1999), interval PLS (Nørgaard et al., 2000), significance tests of model parameters (Westad and Martens, 2000) and the use of genetic algorithms (Leardi and Lupianez-Gonzalez, 1998).

In the following paragraphs is reported a deepening of the only methods of features selection used in this thesis work.

Variable Importance in Projection

In general, a spectroscopic dataset is composed by hundreds of variables, corresponding to the absorbance values at selected wavelengths. Most of the chemical information of the sample is shared within very correlated variables. The PLS algorithm permits to identified the variables more important for regression purposes, furnishing specific plots named Variable Importance in Projection (VIP) plots (Andersen and Bro, 2010; Eriksson et al., 2006; Wold et al., 2001). The VIP scores estimate the importance of each variable in the projection used in a PLS model; the "greater than one" rule is generally used to determine whether a certain variable is actually significant; in this manner is possible to identify the portions of the spectrum that are most useful to calibration. The VIP definition is a combined measure of how much a variable contributes to describe the two sets of data; the dependent (response Y) and the independent variables (descriptor X).

In PLS regression modelling, a variable x_k may be important for the modelling of Y. Such variables are identified by large PLS-regression coefficients, $\mathbf{b}_{mk}$. However, a variable may

also be important for the modelling of X, which is identified by large loadings, $\mathbf{p}_{ak}$. A summary of the importance of an X-variable for both Y and X is given by VIP_k (k variable importance for the projection). This is a weighted sum of squares of the PLS-weights, w'_{ak} with the weights calculated from the amount of Y variance of each PLS component.

VIP is the sum over all model dimensions of the contribution variable influence (VIN). After the definition of the maximum PLS model dimensions **A** (**A** LVs equal to X variable) in comparison to a given PLS dimension **a** (**a** LVs selected) model the parameter called $(VIN)^2_{ak}$ can be calculated. Indeed $(VIN)^2_{ak}$ is equal to the squared PLS weights w_{ak} related to **a** LV selected, multiplied for the explained variance SS of the same LV (PLS dimension). The accumulate values (determined all over PLS dimensions) become:

$$VIP_{ak} = \sum_a (VIN)_k \qquad (3.25)$$

VIP_{ak} is divided by the total explained variance SS and multiplied by the numbers of terms in PLS model. The final VIP_{Ak} can be calculated as the squared root of that values.

The final VIP_{Ak} value for the k^{th} variable is given as the equation 3.28.

$$VIP_{Ak} = \sqrt{\left\{\frac{\sum_{a=1}^{A}\left[w_{ak}^2 \ (SSY_{a-1} - SSY_a)\right] * K}{(SSY_0 - SSY_A)}\right\}} \qquad (3.26)$$

Where w_{ak} is the weight value for variable k in component **a**, SSY_a is the sum of squares of explained variance for the a^{th} PLS latent variable model with **a** selected LVs. SSY_A is the total sum of squares Y explained variance and A is the total number of possible Latent Variables (LVs). SSY_0 represents the original variance of Y matrix.

The average of all calculated VIPs is equal to 1, so a comparison between VIPs is possible. That means a comparison of X variables important for regression purpose. Considering FT-NIR spectra the variables range expressed by the VIPs may be related to the main absorption regions.

The weights in a PLS model reflect the covariance between the independent and dependent variables and the inclusion of the weights is what allows VIP to reflect not only how well the dependent variable is described but also how important that information is for the model of the independent variables. A VIP smaller than one indicates a non-important variable, which could probably be removed.

VIP values are only useful when the overall model is valid and reasonable. Moreover the VIP definition can be applied repeatedly. This procedure implies that the method can be used looking at the original x variables with a critical point of view. The first approach consisted in

highlight only ‘good’ variable intervals (higher than threshold limit equal to 1) and use them for calibration purpose. Otherwise the VIP definition may highlight the ‘bad’ variables instead and remove a few of those (below the threshold equal to 1) before re-running PLS regression method. In both of the cases an improvement in model performances is desirable according to the variable selection model.

It is not advisable to simply remove everything below one. Instead, a few of the variables with the very lowest VIP values should be removed. If the model improves, the method can be repeated on the reduced data set until no more improvements are found.

Interval PLS and Interval PLS-DA

Concerning highly correlated data, such as spectral data, variable selection can be carried out using windows of variables instead of evaluate the importance of each variable individually. One of the most commonly used method is called Interval Partial Least Square (iPLS) (Nørgaard et al., 2000). The iPLS can be used for regression purposes (iPLS) or for classification purposes (iPLS-DA), depending on the response **Y** used.

More in detail, iPLS is an interactive extension of PLS, which develops local PLS models on equidistant subintervals of the full spectrum region. For this reason, the spectra are divided into a number of non-overlapping intervals of equal length (width) and PLS model are made on each of these intervals. The purpose is to search one or more intervals giving the best prediction models in comparison to the prediction model calculated on the full spectrum. The main advantage of using iPLS is the graphical output giving an overview of the spectral data and in displaying interesting spectral areas which could be selected. The corresponding important and selected interval are presented in the informative iPLS plots, in which the original spectra shape is presented (Andersen and Bro, 2010; Leardi and Nørgaard, 2004).

The comparison of interval performances is based on the minimum RMSECV, the corresponding choice of the number of LVs used to build the PLS model and/or the correlation between measured and predicted values. To find the best model it is proper to test intervals of different width and to interpret any significant difference in the results. An iPLS model based on narrow intervals generally needs long calculation time; on the contrary, a model based on wide intervals may have the risk to hindering the effect of small peaks to be seen. In addition to the interval width (or the number of intervals), also the procedure used to include the intervals in the model can be chosen by the user. In fact, iPLS may be used both in

the *forward* and in the *reverse*, or *backward*, mode (FiPLS / BiPLS) (Xiaobo et al., 2010; Balabin and Smirnov, 2011).

More in detail, *forward* iPLS is conceived to calculate local PLS models on each subinterval, then to choose the best one on the basis of the lowest RMSECV value. In the second cycle, the first selected interval is used in all models but is combined with each of the remaining intervals one at a time, and the best combination of the two intervals is chosen again on the basis of the lowest RMSECV value. This iterative procedure is repeated until no further decrease of RMSECV is achieved. The *reverse* iPLS, on the contrary, works by initially including all the intervals in the model, then by discarding a single interval at a time. When discarding a certain interval produces the lowest RMSECV value, that interval is definitively excluded from the model. The same procedure is repeated by discarding the second "worst" interval and so on until no further decrease of the RMSECV values is obtained.

Wavelet Packet Transform for Efficient pattern Recognition

When dealing with signals, the variables constitute a discrete sequence of points: the information of interest may lie not only in the single points, but also in particular aspects of the signals shape. Therefore, the order of the variables should also be taken into account. For these reasons, in this thesis, an algorithms operating the variables selection step in the wavelet domain has been used (Walczak, 2000; Misiti et al., 1999). In fact, the Wavelet Transform (WT) constitutes a very powerful tool, since it maps the signal as a function both of frequency (scale) and of the original domain (Ulrici et al., 2008; Cocchi et al., 2005; Jetter et al., 2000). Therefore, WT allows modeling both aspects of an instrumental signal, i.e., point and shape characteristics. In other words, the representation in the wavelet domain offers the possibility not only to use the single intensity values of the signal, but also peak widths, slopes of selected portions of it, degree of smoothness, more or less evident discontinuities, and many other shape aspects in order to achieve the information sought. Moreover, removing baseline effects and reducing noise is much easier in the WT domain than in the original one, avoiding the usage of spectra pretreatments such as SNV or MSC (Hindle et al., 1996). Finally the degree of compression achieved by WT is extremely high, usually very few wavelet coefficients contain all the information of interest. Besides these coefficients cover contiguous signal regions.

The WPTER (Wavelet Packet Transform for Efficient pattern Recognition) algorithm (Cocchi et al., 2001) is based on the Wavelet Packet Transform (WPT). Given a [m x n] matrix composed by m objects or signals of length n, WPTER (Figure 3.6.1) first checks if the length of the signals equals a power of 2. If they do not, the length n', equal to the power of 2 immediately higher than n is reached by adding a suitable number of zeros, by means of the so-called *zero padding* procedure. All the signals are individually decomposed into the WPT domain until a user defined maximum decomposition level, L, and a three dimensional [$m \times n' \times L$] array is obtained, called *WPMAT* (n' depending on the decomposition level, L), that represents each one of the m considered signals in the Wavelet Packet domain. This is a redundant representation, since one can obtain a perfect reconstruction of the signals in many different ways, i.e., using many different bases, which correspond to various combinations of the blocks.

Before choosing the best basis, a pre-selection of the most informative wavelet coefficients is performed on every approximations and details vector. For each scale $1 \leq s \leq L$ and position $1 \leq p \leq n'$, a discriminant parameter $\alpha_{s,p}$ is calculated over all the m objects, which can be either the (between-class variance)/(mean intra-class variance) ratio or the (total variance)/(mean intra-class variance) ratio of the corresponding wavelet coefficients. This parameter is then used to eliminate all the wavelet coefficients that are located in those positions p and scales s corresponding to the $\alpha_{s,p}$ lowest values. For every block, only a given percentage (specified by the user) of those wavelet coefficients that lead to an efficient distinction of the objects belonging to different classes is retained. Therefore, the best basis has to be selected.

The selection of the best basis requires the choice of an appropriate method to evaluate how efficiently a basis separates the diverse pieces of information contained in the signal. Usually the methods for best basis selection use the entropy content (Coifman and Wickerhauser, 1992), i.e., codify how the energy of the signal is distributed over the coordinates of a given data space. A basis in which only few coefficients attain high values is characterized by a low entropy value, while a basis whose coefficients assume similar values, i.e., a basis in which the information is spread over several coefficients, possesses high entropy. Therefore, it is possible to search for the best basis by minimizing a proper function for measuring entropy. Various ways to quantitatively determine the entropy, e.g. *Shannon* and *log Energy* criteria, are implemented in WPTER. However, using the entropy-based methods, the best basis is selected by looking for the representation of the data set in which the information of interest is concentrated in the smallest possible number of elements, thus

not considering explicitly how the different possible bases lead to effective separation of the individual signals into relevant classes. These considerations suggested to use an alternative procedure to evaluate the discriminant capability of each block deriving from the WPT decomposition. This new procedure, the Classification Ability (CA) criterion, is based on the estimation of the Euclidean distance between each couple of objects (signals) in the thresholded wavelet coefficients space. CA is defined as to attain low values in correspondence of best separation among the objects (spectra) belonging to different classes and, at the same time, best clustering among the objects belonging to the same class. The best discriminant basis is therefore identified as the one containing the approximations and details vectors attaining the lowest CA values. The basis is not forced to be complete, since the goal is not the perfect signal reconstruction, but the identification of those features, which are important to the classification task.

Once the best discriminant basis, called Classification Ability Basis (CAB), has been identified, the wavelet coefficients herein contained can be used for signals reconstruction and classification. Each signal is in fact reconstructed back into the original domain by using only the previously selected wavelet coefficients belonging to the CAB and setting to zero the others. These reconstructed signals represent the projection of the selected wavelet coefficients into the original domain, pointing out the portions of the signals responsible for classification and giving information about the scales at which the features of interest are located. Since in many cases the reconstructed signals are quite different from the original ones, a representation of the mean original signals is also given, where the contiguous regions containing the features of interest are highlighted. These signals representations allow us to give a chemical interpretation of the classification results.

The classification on the basis of the selected wavelet coefficients can then be performed using two methods:

- the Percentage of Assignation (PA) calculated for each one of the reconstructed signals with respect to the mean reconstructed signal of each class, that was the only method implemented in the former version of the algorithm. The PA parameter assumes values varying in the 0-100 range, and it is defined according to the following: each one of the p nonzero points of a given reconstructed signal lying in the interval given by the value of the corresponding point in the mean reconstructed signal of the considered class ± twice its standard deviation contributes with a value of $1 / p$ %. Otherwise, if the considered point does not lie in this interval, its contribution to the corresponding PA value is null. In the ideal event in which every reconstructed signal is assigned in all its points exclusively to the proper class,

PA results equal to 100 for all the signals with respect to their own classes and to 0 for all the signals with respect to the other classes. Obviously, all the intermediate cases are possible. The PA values calculated for each object with respect to each class are represented in a three-dimensional bar graph;

- the SIMCA method applied to the selected wavelet coefficients, that has been implemented in the latest versions of the algorithm.

The classification model created on the basis of the training set signals can then be validated by applying it to a set of test signals. Following the same procedure as for the training set, the test set signals are first decomposed into the WPT domain, and then reconstructed back into the original domain, using only the previously selected wavelet coefficients belonging to the CAB. The reconstructed test set signals and the calculated wavelet coefficients are then evaluated with respect to the existing classes by using the PA and SIMCA methods, respectively. Furthermore, for interpretative purposes, for each class the corresponding mean original signal is plotted, highlighting the regions corresponding to the selected features.

In order to find the optimal classification models it is possible to vary different parameters, such as the type of wavelet used for the WPT decomposition, the maximum level of decomposition, and the percentage of coefficients retained in the thresholding operation. It is -therefore- clear that many trials have to be performed, corresponding to different values for the above cited parameters, before finding the optimal overall conditions for classification. This is made, automatically, in a massive way by WPTER: the calculations are repeated by cycling over all the possible combinations of the chosen parameters. For each cycle, the efficiency of the classification obtained is evaluated by means of a Score Function (SF) based on the PA and the SIMCA Reduced Distance (RD) parameters. The results are shown in score graphs reporting the SF values of each cycle of calculation, where the most effective classification models are easily identified as those giving the lowest values. For each cycle of calculation, the effectiveness of the classification has been also summarised by means of the SENS (percentage of objects of each class accepted by the class model) and SPEC (percentage of objects of the other classes correctly rejected by the class model) values of the corresponding models calculated on the selected wavelet coefficients. The assignment of an object to a class was decided on the basis of the reduced distance from the class model for the SIMCA models and comparing the PA value with the threshold value of 50% for the PA models. If test signals are available, the SF score graph and the SENS and SPEC values for the test signals are also displayed.

In addition to the classification methods already implemented in WPTER, it is possible to use other algorithms, also based on latent variables, to obtain more efficient and robust classification models. In fact, the wavelet coefficients selected in a WPTER model can be appropriately scaled and used as a new set of variables.

In this thesis work, the matrices of wavelet coefficients of the calculated WPTER models were used as input data for obtaining PLS-DA models. This procedure permits to overcome a present lack of WPTER, that does not allow to perform both cross-validation and external validation.

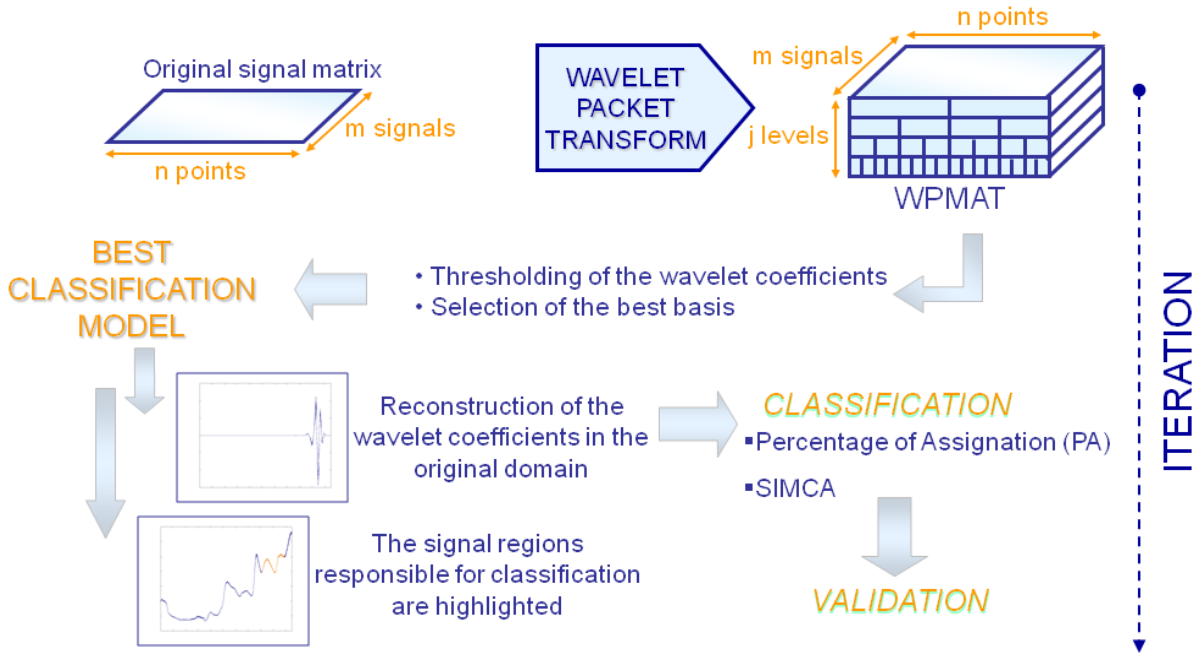


Figure 3.6.1. Schematic representation of the WPTER algorithm steps.

3.7 Software

In the present Ph.D. thesis all calculations were carried out using MATLAB© 7.11. (R2010b). PCA, PLS and PLS-DA and the interval version (iPLS and iPLS-DA) analyses were run using the PLS Toolbox ver. 6.51 for MATLAB© ver. 7.11. (R2010b). (Wise, 2007). WPTER has been written in MATLAB© ver. 7.11 (R2010b) language and use some routines from the Wavelet Toolbox ver. 2.1 for MATLAB© (Misiti et al., 1999) and from the PLS Toolbox for MATLAB© (Wise, 2007).

3.8 References

Andersen, C. M., and Bro, R. (2010). Variable selection in regression – a tutorial. *J. Chemom.* 24, 728-737.

Anderson, T.W., (1984). An Introduction to Multivariate Statistical Analysis, Wiley, New York,.

Balabin, R.M., Smirnov, S.V. (2011). Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, *Anal. Chim. Acta.* 692, 63-72.

Barker, M., Rayens, W. (2003). Partial Least Squares for Discrimination, *J. Chemom*. 17(3), 166-173.

Barnes, R. J., Dhanoa, M. S., Lister, S. J., (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* 43, 772-777.

Centner, V., Massart, D.L., De Noord, O.E., de Jong, S., Vandeginste, B.M., Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* 68, 3851-3858.

Cocchi, M., Seeber, R., Ulrici, A. (2001). WPTER: wavelet packet transform for efficient pattern recognition of signals. *Chemom. Intell. Lab. Sys.* 57, 97-119.

Cocchi, M., Corbellini, M., Foca, G., Lucisano, M.,. Ambrogina Pagani, M., Tassi, L., Ulrici, A., (2005). Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra. *Anal. Chim. Acta.* 544, 100-107.

Coifman, R.R., Wickerhauser, M.V.,(1992). Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory.* 38, 713-718.

Eriksson, L., Johansson, E., Kettaneh-Word, N., Trygg, J., Wikstrom, C., and Word, S., (2006). Multi-and Megavariate Data Analysis Part I and Part II Umetric academy.

Forina, M., Oliveri, P., Jäger, H., Römisch, U., & Smeyers-Verbeke, J. (2009). Class modeling techniques in the control of the geographical origin of wines. *Chemom. Intell. Lab. Syst.* 99, 127-137.

Forina, M., Casolino, C., Pizarro-Millan, C., (1999). PLS: a technique for the elimination of useless predictors in regression problems. *J. Chemom.* 13, 165-184.

Fearn, T. (2009). The effect of spectral pre-treatments on interpretation. *NIR news.* 20, 16-17.

Geladi, P., (2002). Some recent trends in the calibration literature. *Chemom. Intell. Lab. Sys.* 60, 211-224.

Geladi, P., MacDougall, D., and Martens, H., (1985). Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Appl. Spectrosc.* 39, 491-500.

Geladi, P., Kowalski, B.R., (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta*.185, 1-17.

Guo, Q., Wu, W,. Massart, D.L., (1999). The robust normal variate transform for pattern recognition with near-infrared data. *Anal. Chim. Acta.* 382, 87-103.

Hindle, P.H., Smith, C.R.R., (1996). A comparison of calibration robustness relating to different data treatments of a standard set of spectra. *J. Near Infrared Spectrosc.* 4, 119-128.

Isaksson, T., and Naes, T., (1988). The effect of Moltiplicative Scatter Correction and Linearity Improvement on NIR spectroscophy. *Appl. Spectrosc.* 42, 1273-1284.

Jackson, J.E., (1991). A users guide to principal components, Ed. Wiley & Sons Ltd., Chichester.

Jetter, H., Depczynsky, U., Molt, K., Niemöller, A. (2000). Principles and applications of wavelet transformation to chemometrics. *Anal. Chim. Acta*. 420, 169-180.

Leardi, R., and Norgaard, L., (2004). Sequential application of backward interval PLS and genetic algorithms for the selection of relevant spectral regions. *J. Chemom.* 18 (11), 486-497.

Leardi, R., Lupianez-Gonzalez, A. (1998). Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.* 41, 195-207.

Lindgren, F., Geladi, P., Rannar, S., Wold, S. (1994). A PLS kernel algorithm for data sets with many variables and fewer objects. *J. Chemom.* 8, 111-125.

Todeschini, R., (1998). Introduzione alla chemiometria. EdiSES,

Martens, H., (2001). Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression. *Chemom. Intell. Lab. Syst.* 58, 85-95.

Martens, H., Naes, T., (1989). Multivariate Calibration, Ed. Wiley & Sons Ltd., Chichester.

Massart D.L., Vandeginste B.G.M.,. Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J., in Vandeginste, B.G.M.; Rutan, S.C. (Eds.), (1997). Handbook of chemometrics and qualimetrics: Parts A and B, Elsevier, Amsterdam.

Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.M., (1999). *Wavelet Toolbox User's Guide*. MathWorks, Natick, MA, USA.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B. (2000). Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl. Spectrosc.* 54, 413-420.

Rinnan, A., Van den Berg, F., Engelsen, S.B. (2009). Review of the most common pre-processing techniques for near-infrared spectra, *Trends Anal. Chem.* 28, 1201-1222.

Savitzky, A., and Golay, M.J.E., (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Anal. Chem.* 36, 1627.

Stone, M., (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statist. Soc. B* 36, 111-147.

Ulrici, A., Cocchi, M., Foca, G., Durante, C., Marchetti, A., Tassi, L. (2008). New Trends in Analytical, Environmental and Cultural Heritage Chemistry: Cap. 5 Multivariate analysis of analytical signals to decipher relevant chemical information. Editors: Colombini, M.P., and Tassi, L.

Walczak, B., (2000). Wavelets in Chemistry, Elsevier, Amsterdam, NL.

Westad, F., Martens, H., (2000). Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression. *J. NIRS* 8, 117-124.

Wise, B.M., Gallagher, N.B., Bro, R., Shaver, J.M., Windig, W., Scott Koch, R. (2007). *PLS toolbox 4.2*. Wenatchee, WA: Eigenvector Research Inc.

Wold, S., Sjostrom, M., Eriksson, L., (2001). PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.* 58, 109-130.

Wold, S., Esbensen, K., Geladi, P., (1987). Principal component analysis. *Chemom. Intell. Lab. Syst.* 2, 37-52.

Workman, J. Jr., (2002). The state of multivariate thinking for scientists in industry: 1980-2000. *Chemom. Intell. Lab. Syst.* 60, 13-23.

Xiaobo, Z., Jiewen, Z., Povey, M.J., Holmes, M., Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. A Review. *Anal. Chim. Acta.* 667, 14-32.

Chapter 4: MULTIVARIATE CALIBRATION MODELS OF WHEAT PARAMETERS RELATED TO QUALITY

4.1 Introduction

This part of the thesis is devoted to studying the possibility to use NIR spectroscopy for the prediction of quality parameters of wheat samples in different physical forms, used in the production of baked products. Indeed the great variety of different baked products in Italy has led to the development of a method, named the *Synthetic Index of Quality* (ISQ), for the classification of bread wheats, destined to different end uses, in different quality categories. Based on chemical and rheological measurements, each wheat sample is normally assigned to the most suitable class by an expert assessor. Often this procedure makes the class assignation uncertain, thus leading to the possibility of controversies during the trading phase.

Most of the ISQ parameters require the execution of time demanding and expensive chemical and rheological laboratory measurements. Multivariate calibration models based on NIR spectra could be therefore very useful in order to determine at the same time the value of the various wheat quality-related parameters (i.e., protein content, hectolitre weight, alveograph W and P/L, farinograph stability, falling number, sodium dodecyl sulphate (SDS) sedimentation index) without the need of sample pretreatment.

For this reason, the present research is aimed at building calibration models able to predict the values of the ISQ parameters, in addition to the SDS index, for different flours starting from NIR spectra. To this aim, different calibration approaches have been used.

4.2 Wheat

Origin and morphology

Wheat is a wild grass native to the 'fertile crescent' region. Since the 8000 B. C., this grass represented the main foodstuff for the societies located in Mesopotamia and Egypt. The favourable climate, the location and the large amount of seeds that can be obtained from these plants were the principal reasons that transformed the hunters/gatherers societies into stable populations characterized by permanent villages that thrived by the start of the modern agriculture and became the first case of ancients civilizations from the historical point of view.

Nowadays, more than 4000 modern wheat varieties grown around the world. These varieties are the results of progressive and gradually selection and modification made by

farmers during the centuries in a way to obtain improved foodstuffs. It also became the leading grain used for human consumption due to its nutritive profile and relatively easy harvesting, storing, transportation and processing, as compared to other grains and cereals.

A longitudinal and cross section of a wheat caryopsis (or kernel) along with an identification of its components is shown in Figure 4.2.1 (Le Scienze Blog: italian edition of Scientific American; Cappelli et al., 2000). Independently of plant cultivar *Triticum Aestivum* or *Triticum Durum* (respectively called "soft" wheat and "durum" wheat), the morphology of the wheat caryopsis is unique and essentially consists in the germ, the endosperm and some covering surfaces layers (the outer one is called pericarp).

The separation of the endosperm and the germ from the outer fibrous layers, commonly named “bran” creates some technical challenges during the milling stage. The wheat germ (about 2-4% of the caryopsis weight) is located on the dorsal side. The wheat germ parts are the embryo, with rudimentary roots and shoots, and the scutellum, which is a transport organ of nutrition to the embryo during sprouting. The wheat caryopsis outer botanical coats (about 7-8% of the caryopsis weight) consist of several distinct cellulose-rich layers. The outermost layer, the pericarp (fruit coat), is made up of the outer pericarp, which includes the outer epidermis, hypodermis, thin-walled cells, and the inner pericarp, which includes intermediate-size cells, cross-layers, and tube cells (inner epidermis). The inner layers are the seed coat (head) and nucellar epidermis (hyaline layer). Between the nucellar epidermis and the starchy endosperm we find the aleurone layer, having high soluble protein and mineral contents. The aleurone layer constitutes about 5-8% of the wheat kernel. This layer is botanically similar to the endosperm, but it is difficult to separate from the bran by conventional milling techniques. Depending on the kind of wheat, the thickness of the aleurone layer varies. Mechanical damage or hydrolysis with cellulase of the aleurone thick cell wall allows access to protein within the aleurone layer. Although nutritious, incorporation of a fraction with a large percentage of aleurone layer adversely affects the baking quality of flour. The endosperm of the caryopsis was also shown to follow a gradient in ash, protein, gluten characteristics, and baking quality, going from its outer to its inner part.

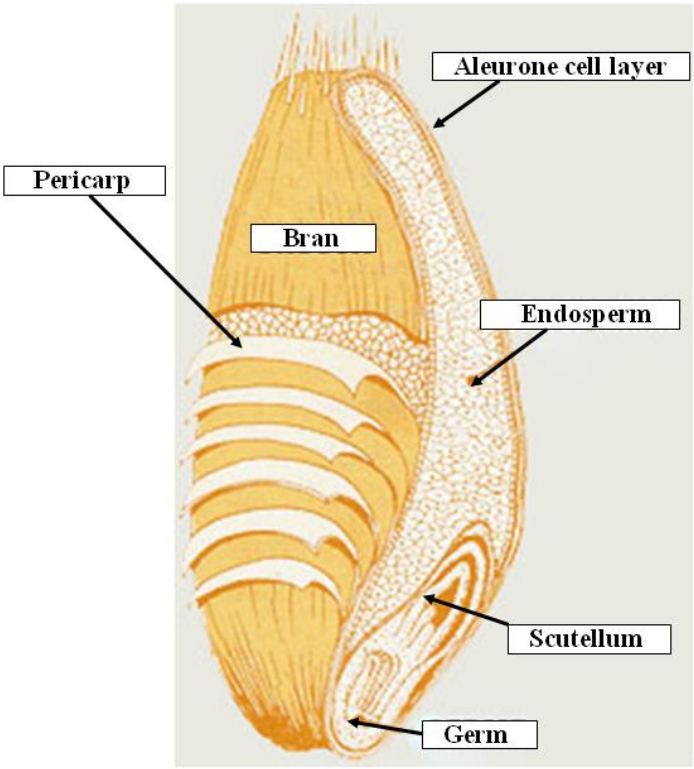


Figure 4.2.1. Cross section of the wheat caryopsis.

Wheat constituents

The major components of the wheat caryopsis (Posner, 2000) are summarized in Table 4.2.1. Among them, the most important one for the technological characteristics of the industrial and homemade bakery products is the protein content. The wheat protein fraction is constituted by insoluble proteins (albumins), proteins soluble in saline solutions (globulins), proteins soluble in alcohol (prolamines) and proteins soluble in acid or basic solutions (glutelin). The different proteins are distributed in the bulk of the kernel in a inhomogeneous manner. In fact, the proteins that are present in the inner zone of the kernel are prolamins and glutelins, about in the same quantity. The germ proteins are particularly albumins and globulins, while in the bran are more abundant the prolamins, accompanied by little quantities of albumins and globulins (Fennema, 1996).

	Caryopsis (%)	Embryo (%)	Scutellum (%)	Pericarp (%)	Aleurone (%)	Endosperm (%)
Protein	12.6	35.0	26.0	5-7	18	7.4-13.9
Ash	1.9	5.45	5.0-8.0	3-4	14-17	0.28-0.40
Fat	1.6	16.3	32.0	3-5	10	0.8-1.5
Starch	59.2	-	-	-	-	68
Pentosan	6.7	6.6	6.6	34.9	39.0	1.4
Cellulose	2.3	2.0	2.0	38.4	3.5	0.3

Table 4.2.1. Average values of constituents of wheat with 14% of moisture.

The wheat flour presents a particular protein composition, that makes it more proper than other cereals for bread and pasta making. The protein fraction of the flour, about 7-15%, is of two types. One type (about the 15% of the total) consists of residues of the typical cytoplasmatic proteins, mostly enzymes, which are soluble in water or dilute salt solutions. The remaining 85% are proteins of the seed principally dedicated to the storage of nitrogen, insoluble in ordinary aqueous media and partially insoluble in salted solution such as Sodium Dodecyl Sulphate solution (SDS). This fraction is responsible of the dough formation (Fig. 4.2.2.) (Graveland et al., 1982; Pasqualone et al., 2006). These dough-forming proteins are collectively referred to as *gluten*. The gluten can be readily extracted from the flour by adding enough water to form a dough, leaving the dough to stand for half an hour or so, and then finally kneading the dough under a stream of cold water, which washes out all the soluble material and the starch granules. The resulting tough, viscoelastic and sticky material contains about one-third protein and two-thirds water.

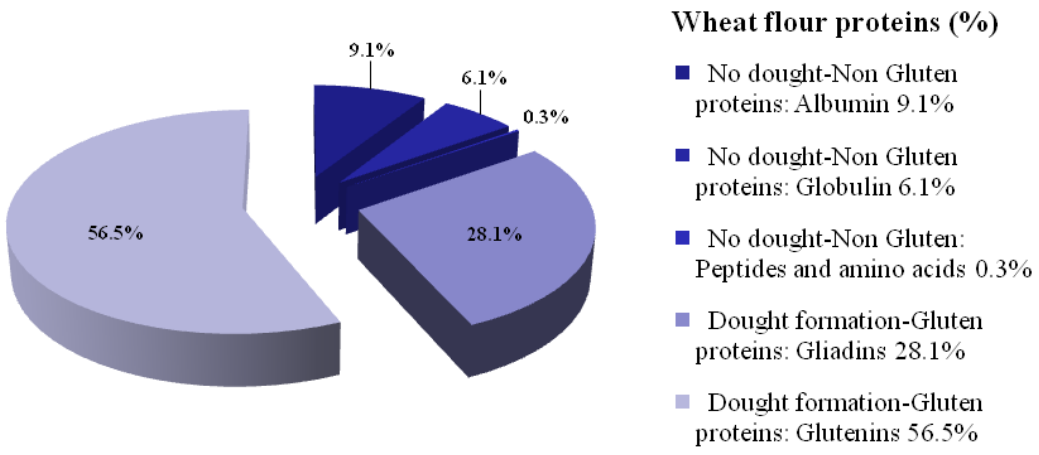


Figure 4.2.2. Proteins of the wheat flour.

The gluten proteins can be fractionated on the basis of their solubilities. The most soluble ones, the gliadins, can be extracted into 70% ethanol. The gliadins constitute about one-third of the gluten. The remaining two-thirds are the glutenins, which are extremely difficult to dissolve fully. One of the most used protein denaturant solution is 1.5% Sodium Dodecyl Sulphate solution (SDS) able to separate gluten proteins from wheat flour (Graveland et al., 1982). It is generally believed that, during dough development, the glutenin proteins molecules are stretched out into linear chains which interact to form elastic sheets under the gas bubbles. A number of chemical reactions are involved in this process (Danno and Honesey, 1982; Borneo and Khan, 1999). The mechanical stresses are sufficient to break, temporarily, the hydrogen bonds that are important in binding together the different gluten proteins. During the early stages of mixing of the dough, the polypeptide chains of both gliadins and glutenins tend to become aligned alongside each other. This gives many more opportunities for hydrogen bond formation, and the resistance of the dough to mixing shows a sharp increase (Figure 4.2.3).

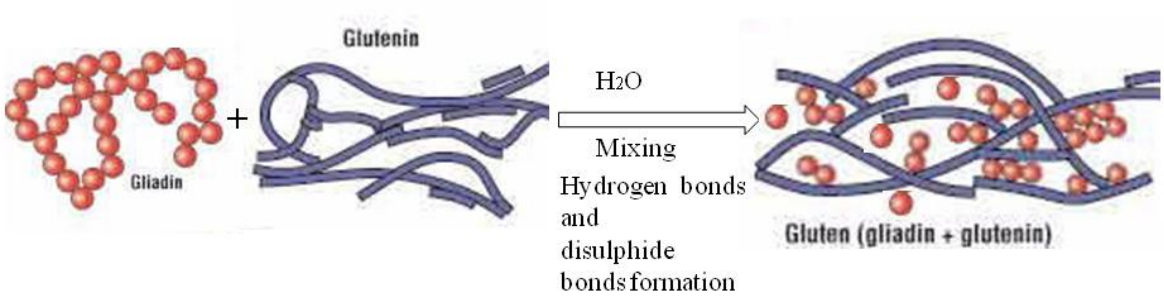


Figure 4.2.3. Dough development (gluten formation) of wheat flour mixed with water and air.

Other reactions involve the sulphydryl groups of the proteins (Schofield et al., 1983). Under mechanical stresses exchange reactions between neighbouring sulphydryl groups will allow the glutenin subunits to take up more extended arrangements, leading to a particular conformation in the secondary structure called ‘β-spiral’, due to a high number of ‘β-bend’ structures, arranged in a sequence (Figure 4.2.3). The β-bends are a configuration of a polypeptide chain that occurs in most proteins at the ends of sections of α-helix, where they allow the rigid helix to change direction. There is therefore a lot of support to the idea that the elastic component of dough’s viscoelasticity can be accounted for by the amino acid sequences of its proteins.

Ash content refers to the mineral content of a flour (principally composed by phosphates of Mg and K, and by salts of Ca, Fe, S, Cu and Zn), and is determined by burning a given quantity of flour under prescribed conditions and measuring the residue. The mineral content varies and depends on many factors, such as the variety of wheat, the soil, the fertilization, and the climate (Singh et al., 1998). The greater portion of minerals found in a kernel of wheat is contained in the germ, and husk, or bran, and the least amount in the endosperm (Hinton, 1959; Morris et al., 1945). As a consequence, if a flour contains a greater number of bran particles, it has a more elevated ash content. The determination of the ash content serves to estimate the degree of the endosperm separation from the bran during milling, i.e. the grade of flour. Generally, flours considered to be of higher quality are more refined and produce less ash.

Starch is found in greatest abundance in the endosperm of the seeds (Coultate, 2005). It occurs in the form of granules which have usually an irregular rounded shape, ranging in size from 2 to 100 μm.

Starch consists of two types of glucose polymer: *amylose*, which is essentially linear, and *amylopectin*, which is highly branched (Lehninger, 1975; Miles et al., 1985).

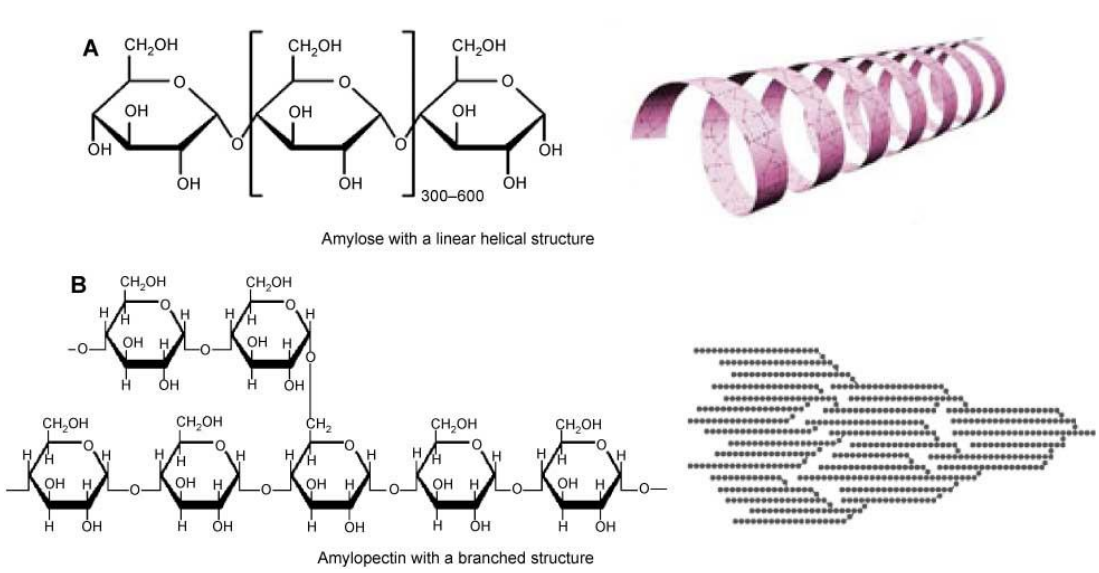


Figure 4.2.4. Chemical structure of amylose (A) and amylopectin (B).

Most starches are 20-25% amylose and 75-80% amylopectin. Amylose consists of long chains of α-D-glucopyranosyl residues linked between their 1- and 4-positions (Fig. 4.2.4A). They average molecular weights are about 2 x 10^6 dalton. Amylopectin is a much larger molecule, having about 10^6 glucose units per molecule. As in amylose the glucose units are

joined by α1→ 4 glycosidic links, but some 4-5% of the glucose units are also involved in α1→ 6 links, creating branch points as shown in Figure 4.2.4B. This proportion of branch points results in an average chain length of 20-25 units. Its essential feature is a skeleton of singly branched chains approximately 40 glucose units long (named B chains), which carry clusters of mostly unbranched chains approximately 15 units long (named A chains).

In the native form of starch these two types of polymer are organized in granules as alternating semi-crystalline and amorphous layers, having in most starches central symmetry (Fig. 4.2.5) (Hermansson and Svegmark, 1996). The semi-crystalline layers consist of ordered regions composed of clusters formed by short amylopectin branches, most of which are further ordered into crystalline structures. The amorphous regions of the semi-crystalline layers and the amorphous layers are composed of amylose and non-ordered amylopectin branches. Moreover, there is an additional complexity relating to the nature of the crystalline structures: the clusters comprising the crystallites are densely packed in an orthogonal pattern that also contains structural water.

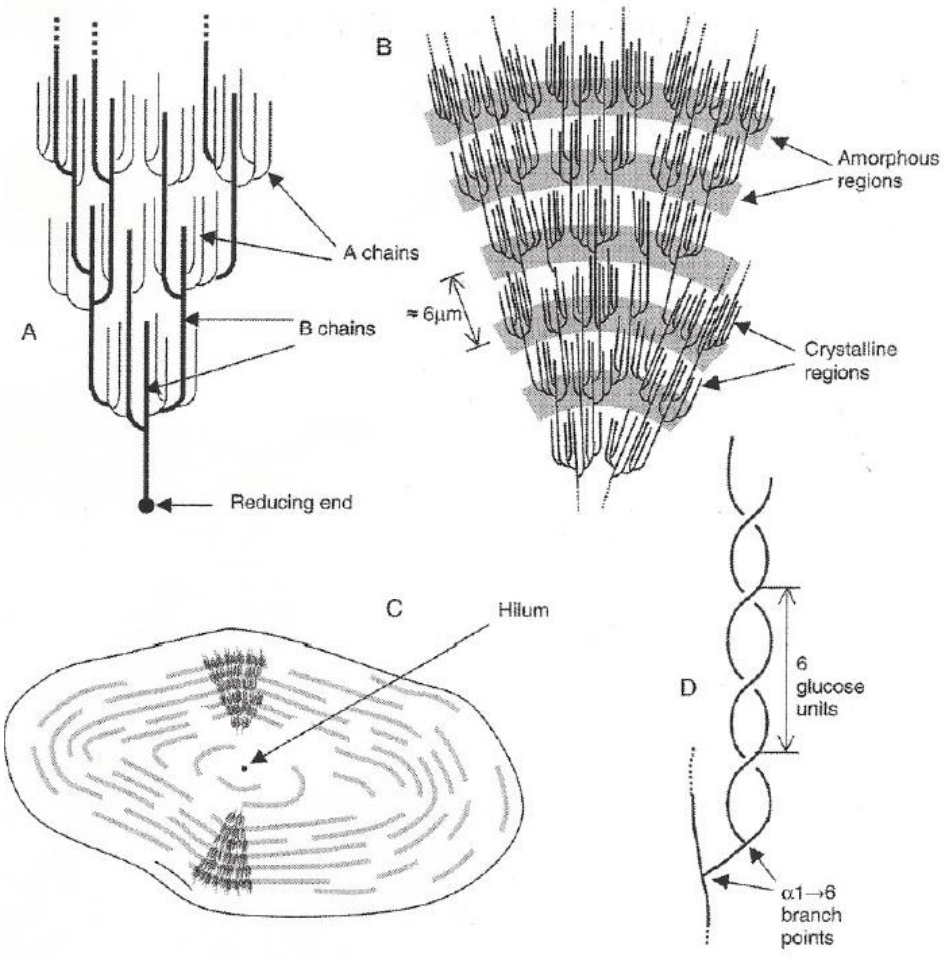


Figure 4.2.5. The clusters of the amylopectin branches forming the semi-crystalline layers.

Cellulose is the major building block of the cell wall structure of higher plants (Coultate, 2005). It is usually associated with the structural components hemicellulose and pectin. Collectively, these three chemical systems are referred to as the plant structural polysaccharides. Cellulose is unbranched, and the chain may contain up to 10.000 β-D-(1→4)-linked glucopyranosyl units (Figure 4.2.6). The extended molecule forms a flat ribbon, which is further stiffened by intra- and intermolecular hydrogen bonds that produce a regular structure with low solubility. Native cellulose is, however, composed of both highly ordered, crystalline, and amorphous regions. Although the polymers are insoluble in water, the amorphous cellulose has the ability to readily absorb water.

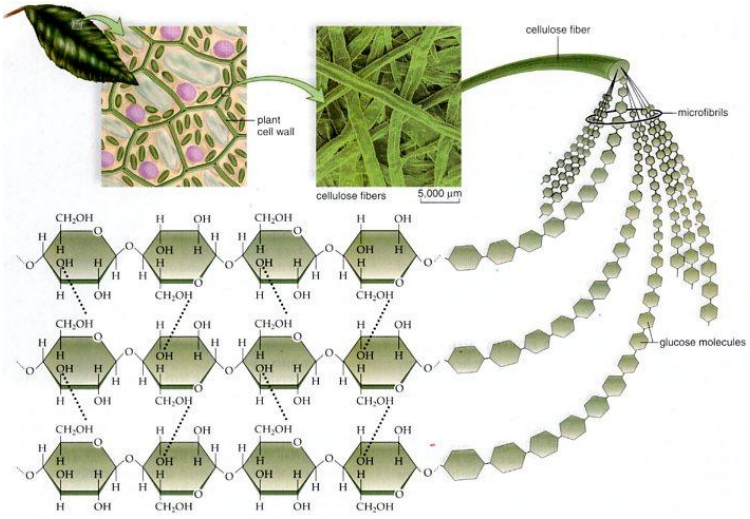


Figure 4.2.6. Chemical structure of cellulose.

Lipids are relatively minor constituents in wheat grains (Posner, 2000). However, they must be taken into account when discussing nutrition, grain storage, processing such as dry and wet milling, brewing, cooking and extrusion. All cereal grain lipids are rich in unsaturated fatty acids (FA). Palmitic acid (16:0) is a major saturated FA, while linoleic acid (18:2) is a major unsaturated FA for most of the cereals.

The milling process

The grinding of the wheat kernels (that are also called *grit*), is a process where the mill have to get the greatest possible amount of flour of the highest possible quality (Palmer, 1989). The concept of "*Quality*" is associated to the special characteristics of the flour in relation to its end use (cookies, bread, pasta, etc.). From a technological point of view, the low sifting of flour produces a sacrifice of the more healthy components of the kernel, present in

the external layers of the caryopsis, such as the proteins of the aleuronic layer. A good percentage of these components are, however, present in more highly sifted flour. The present grinding processes allow to obtain different types of flour, based on the particle size chosen. The chemical characteristics of the flour and the sifting rate are linked in a close manner, although the final characteristics of the flour also depend on the properties of the milled grain (Greer and Ziegler, 1974). A sifting rate of 50% excludes the most external parts of the kernel; passing to flours subjected to a sifting rate of 72% and of 80%, the percentage of the parts closer to the outer layers of the kernel is always greater.

During milling, the wheat germ has to be removed, since it is rich in unsaturated fatty acids that may undergo to oxidation, compromising the preservation of flour. The modern milling industry is mainly concentrated on the inner core of the grain (endosperm), the part of the kernel where the most valuable nutritional substances are stored and where a high concentration of starch and proteins is contained.

The phases of wheat milling (Posner, 2000; Greer and Ziegler, 1974) are summarized in the diagram shown in Figure 4.2.7.

The first stages of milling include some cleaning steps, to prevent the contamination of flour. In modern mills cleaning is performed using dry methods. Cleaning removes both foreign bodies (rocks and their fragments, straw, etc.) and very small and light particles, such as dust, fragments of insects and eggs, etc. During these processes the removal of specific parts of the caryopsis (for example, the beards) also occurs. The cleaning techniques used are quite sophisticated and are based on the exploitation of the characteristics of grain from impurities and foreign objects (dimensions, shapes, differences in specific density, magnetism, etc.).

Grinding is usually preceded by conditioning operations, where temperature and humidity of the grain are optimized before passing through the rolling mill. A suitable moisture of the grains promotes optimal rupture of the grains, and the subsequent separation of the larger portions of the bran from the endosperm. The actual grinding is the result of alternating passages of grain through rolling mills and sieving tools (plansichters and purifiers). The plansichters, which are made up of overlapping oscillating sieves with decreasing mesh size, have the task of separating the ground wheat based on the size of its particles. The product that comes out of plansichters is then conveyed, by means of pneumatic transport devices, to the subsequent stages of the grinding process, that are weighing and storage. Depending on bran amount, different types of flour can be obtained by the milling process; the flour that contains a higher amount of bran is normally called *wholemeal* flour.

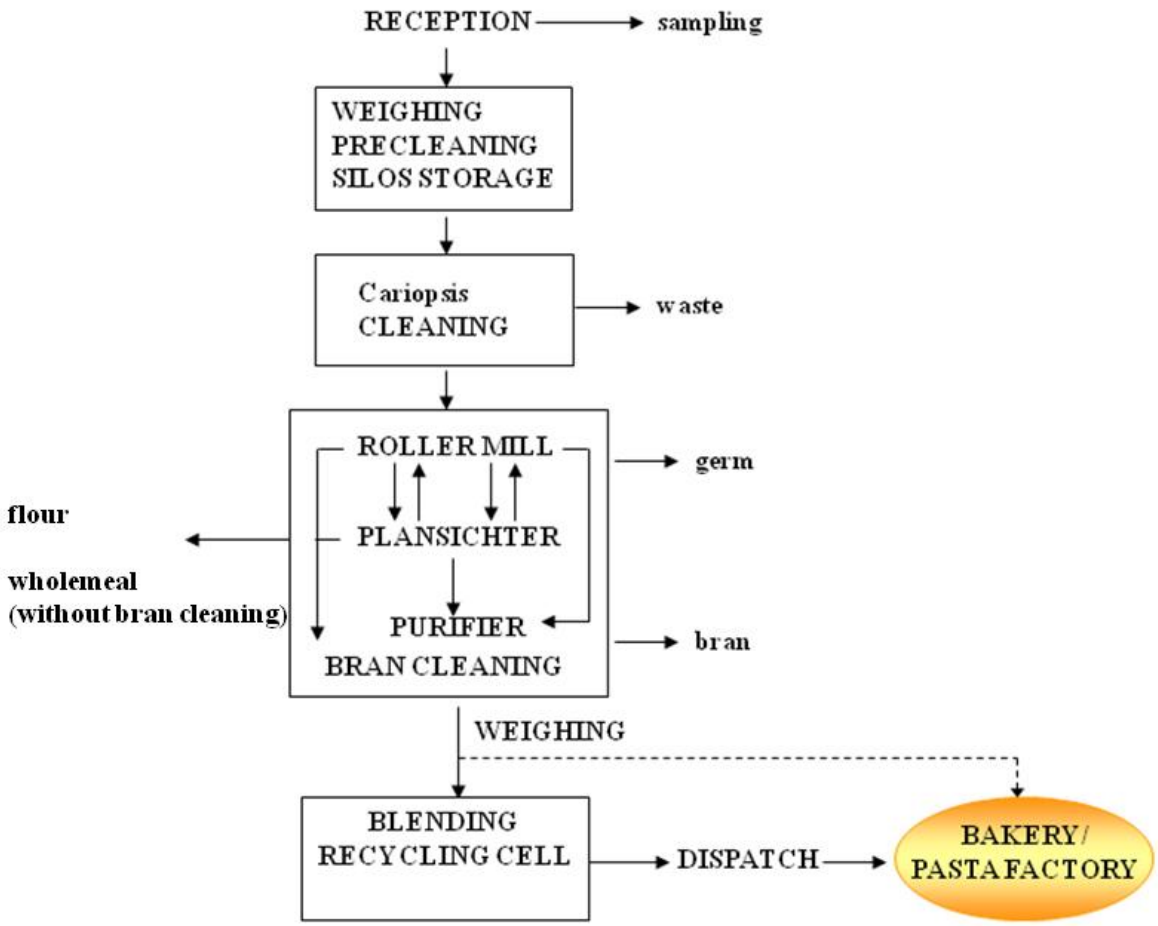


Figure 4.2.7. Phases of the milling process.

The Synthetic Index of Quality

The variability of wheat bread is influenced by many factors arising from the type of variety grown, the farming techniques used in relation to environmental and seasonal effects, as well as from the production area and the climate changes recorded year after year. The quality of the wheat, and consequently its price, is influenced by all these sources of variability and by their interactions. The major risk for both buyers and sellers is that the product will not meet requisites and expectations when delivered. In addition, the milling industry requires huge batches of bread wheat, already characterized for quality. This request responds to two specific requirements:

i. to store various types of wheat separately on the basis of their properties and in relation to the final use and;

ii. to properly mix the different grains in order to obtain the optimum flour for the special processing of interest.

For example, flours with high concentration of gluten proteins are able to withstand high stresses during the technological processes of mixing and leavening during the production of bread, while flours with weaker gluten and a lower amount of proteins are preferred for the production of biscuits and other chemically leavened products. It is therefore clear that the classification of the different lots of wheat is a matter of fundamental importance in the cereal industry.

Since the late 1960s the Italian classification of wheat flour was based on protein, moisture and ash contents of flour, according to the law (Legge 580, 4 luglio 1967). The wheat flours were at the time classified in 5 different trade categories, as reported in Table 4.2.2. (Borasio, 1997). However, this kind of classification furnished only partial information about the technological behaviour of wheat flours during dough-making. Indeed the huge variety of baked products in Italy has led an Italian cereal trade association, ASS.IN.CER., to develop the Synthetic Index of Quality (ISQ) classification method. In particular, wheat samples are classified by means of expert end trained assessors into four different quality categories on the basis of the values of significant parameters, related to the protein content, rheological properties and behaviour in enzymatic tests of the flours (Borasio, 1997; Foca et al., 2007).

	Moisture (%)	**Ash (%)**	**Proteins (%)**
Trade category	**Max**	**Min-Max**	**Min (N x 5.70)**
Bread wheat flour named 00	14.5	0-0.55	9.00
Bread wheat flour named 0	14.5	0-0.65	11.00
Bread wheat flour named 1	14.5	0-0.80	12.00
Bread wheat flour named 2	14.5	0-0.95	12.00
Whole bread wheat flour	14.5	1.30-1.70	12.00

Table 4.2.2. Classification of bread wheat flours according to the Italian law.

The four ISQ categories, shown in Table 4.2.3. along with their corresponding parameters values, are indicated with their Italian names and acronym, going from the stronger to the weaker flour: *Improver Wheat* (FF, Frumento di Forza), *Superior Bread Making Wheat* (FPS, Frumento Panificabile Superiore), *Ordinary Bread Making Wheat* (FP, Frumento Panificabile), *Wheat for Biscuits* (FB, Frumento per Biscotti). When the samples have not the characteristics to be assigned to any class, they are classified as *Flour for Other Uses* (FAU, Farine per Altri Usi). The ISQ categories are exclusively used in the classification of bread

wheat (*Triticum aestivum*) genotypes. Most of the ISQ parameters are represented by traditional chemical analyses and rheological determinations. While some of these parameters are measured in very short time (e.g., hectolitre weight), others require long times for their measurement and are highly liable to measurement errors due to the operator. Furthermore, once the ISQ values are obtained, skilled assessors are required to assign the samples to the proper quality classes. Unfortunately, during the commercial negotiations, the wheat products have to be characterized in very short times.

For these reasons, multivariate calibration models based on NIR spectra could be very useful in order to determine the value of the various wheat quality-related parameters without the need of sample preparation. Nowadays, in the cereal industry, NIR spectroscopy is commonly used with the aim of quantifying different chemical components of the flour, such as moisture and proteins. Moreover, the FT-NIR technique has been successfully employed on wheat flours in order to evaluate particular qualitative parameters of the sample (Cocchi et al., 2005; Foca et al., 2009) and for the prediction of various chemical and technological indexes, which are important to estimate the flour quality (Delwiche et al., 1998; Jirsa et al., 2008). Frequently, some correlations were found between the acquired spectra and some technological properties (Dowell et al., 2006; Sgrulletta and De Stefanis, 1997; Barton II et al., 2000), and the values predicted by the obtained calibration models have been used in order to decide the end-use of the flours.

	Improver Wheat (FF)	**Superior Bread Making Wheat (FPS)**	**Ordinary Bread Making Wheat (FP)**	**Wheat for Biscuits (FB)**	**Points**
Proteins amount (%)	12.5-13.5	10.5-11.5	9.0-10.0	11.0-10.0	70
	13.5-14.5	11.5-12.5	10.0-11.0	10.0-9.0	100
(Prot_ss%)	> 14.5	> 12.5	> 11.0	< 9.0	130
Alveographic W (cm^2)	270-300	220-250	140-170	140-110	70
	300-340	250	170-200	110-80	100
(W)	> 340		> 200	< 80	130
Alveographic P/L ratio	1.8-1.2	1.2-0.8	1.2- 0.7	0.7-0.5	70
	1.2-0.7	< 0.8	< 0.7	< 0.5	100
(P/L)	< 0.7				130
Farinograph Stability (min)	11-13	7-9	3-5	< 4	70
	13-16	9-11	5-6		100
(Stab)	> 16	> 11	> 6		130
Hectolitre Weight (kg/hl) (Phl)	> 75	> 75	> 75	> 75	
Falling Number (second) (FN)	> 250	> 220	> 220	> 220	

Table 4.2.3. ISQ classification of bread wheat flour by ASS.IN.CER. The four quality classes are organized in decreasing order of strength from left to right.

4.3 Materials and methods

Samples

The development of calibration models of general validity requires the use of a set of samples very heterogeneous for genotype, environment and year of harvesting. For this reason, 303 bread wheat samples harvested in two different farming years (153 samples the first year and 150 samples the second year), representative of different bread wheat types, were collected from experimental fields located in different Italian regions.

The samples were stored in three different physical forms: as white flour (_F), as wholemeal flour (_S) and as grit (_G).

Analytical methods for the determination of wheat flour properties

All the wheat samples were characterized by chemical and rheological analyses, consisting in the measurement of seven experimental parameters. Six of them are included in the ISQ quality classification: Falling Number (FN), hectolitre weight (Phl), dry matter protein content (Prot_ss%), alveograph W (W), alveograph P/L ratio (P/L), farinograph stability (Stab). The remaining parameter is the Sodium Dodecyl Sulphate (SDS) sedimentation index (Indrani et al., 2007; Delwiche et al., 1998), that is not part of the ISQ system but it is related to the gluten protein amount and its usefulness is demonstrated in the wheat flour quality definition (Alava et al., 2001).

Every sample has been analysed following the corresponding analytical reference method at the Consiglio per la Ricerca e la sperimentazione in Agricoltura (CRA) in S. Angelo Lodigiano. Once the ISQ analyses have been made, a trained evaluator attributed the 303 samples to the specific classes. The samples resulted to be 24 FAU, 30 FB, 55 FF, 129 FP and 65 FPS. Table 4.3.1. reports the characteristics of the reference methods used in the ISQ classification, in terms of time required for the analysis and number of samples analyzed simultaneously.

The Chopin alveograph used to determine the values of W and P/L ratio is shown in Figure 4.3.1a (Shuey and Tipples, 1982). The alveographic indexes W and P/L are related to strength and extensibility of the dough. A rotating platform shapes a round disc of dough with controlled thickness. The disc is placed on a plate equipped with a device that pushes pressurized air against the disc, forming a bubble that expands until it breaks.

Analysis	Reference method	Instrumentation	Time required for the analysis	N° of samples analysed simultaneously
Proteins	ICC 105/1	Foss Tecator 1002 Distilling Unit	2-3 h	6
Alveographic indexes	ICC 121-1992	Chopin Alveograph PE 87	1 h	1
Farinograph stability	ICC 115-D 1986	Brabender Farinograph SEW	1 h	1
Falling Number	AACC 56-81B 1992	Perten FN 1500	15 min	1

Table 4.3.1. Reference methods and corresponding required needs for the determination of ISQ compositional and rheological indices of the flours.

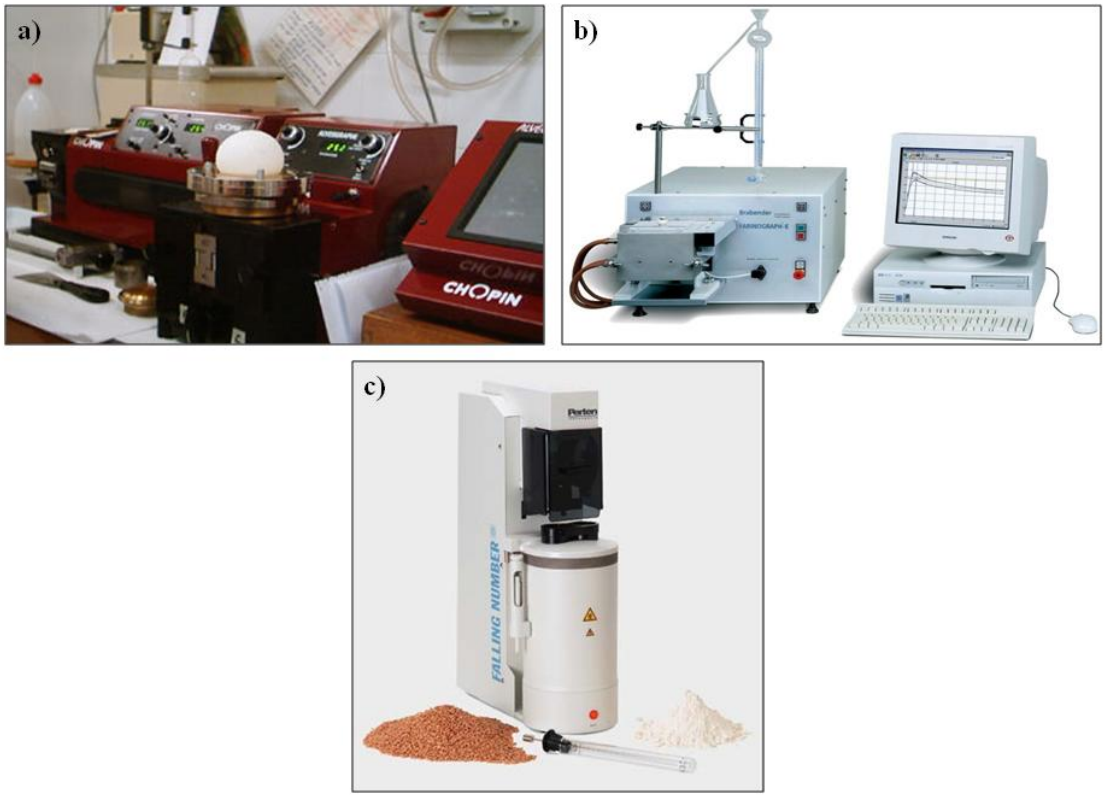


Figure 4.3.1. The Chopin alveograph (a), the Brabender farinograph (b) and the Falling Number measuring instrument (c).

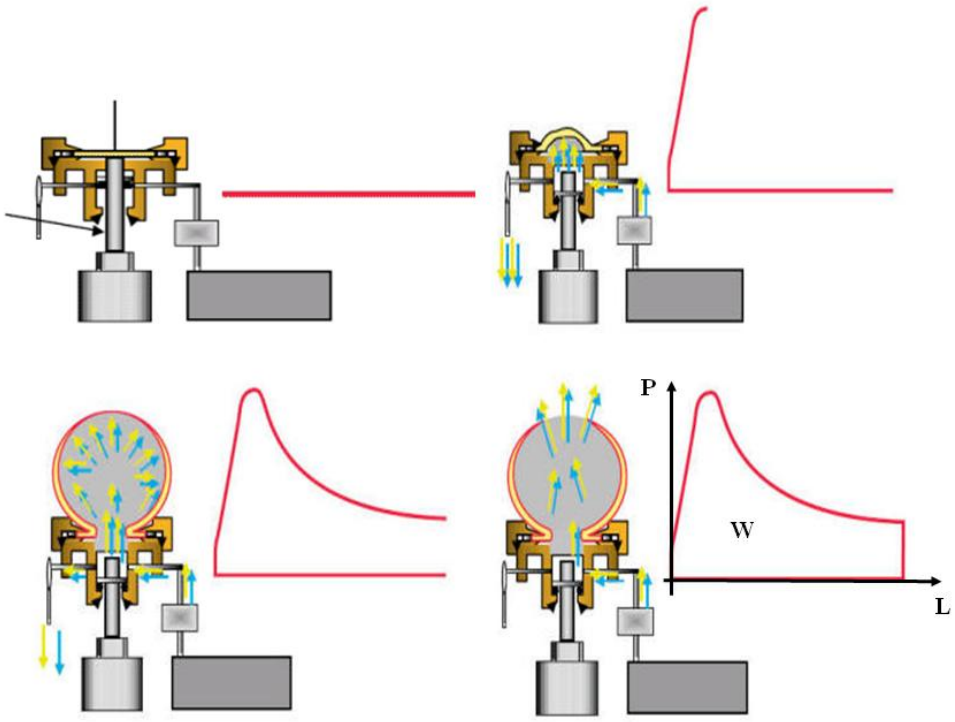


Figure 4.3.2. The creation of the plot by the Chopin alveograph.

The increase in linear dimension and volume of the bubble of dough during the measurement is recorded on a roll of graph paper as elongation (L) as a function of pressure (P) of air applied to the disk. The graph obtained, called alveogram (Figure 4.3.2.), provides three alveographic values related to the dough: resistance to elongation, i.e., the maximum height reached by the curve (P), extensibility, i.e., the total length (L) from the moment in which the disc begins to blow up until the rupture of the bubble, and strength of the dough, indicated by a "W" (cm^2), corresponding to the area under the alveogram.

The higher is the value of "W", the greater the reference area will be; the "W" value indicates the dough strength. Because the maximum height of the curve is related to the total protein content of the flour, the P/L ratio of the Chopin alveograph is a very important information on the gluten content and the dough quality of the flour being tested.

The Brabender farinograph allows to determine the value of Stability (Stab) present in the ISQ classification method. The corresponding instrumental apparatus is shown in Figure 4.3.1b. (D'Apollonia and Kunerth, 1984). The stability of a dough is related to its consistency and to the quantity of water it requires. The method consists in the mixing of 50g of flour, for 20 minutes, with the quantity of water necessary to reach the optimal consistency of 500 Brabender Units (B. U.). The principle on which the farinograph is based is to refer to the resistance of the dough to a certain mechanical stress applied by keeping constant the components causing the stress (e.g. the mixing speed) as well as the experimental conditions

in which this stress is applied (e.g. temperature). The Brabender farinograph mixes a defined quantity of dough and simultaneously plots the graph on a roll of graph paper, representing on the x-axis the time and the y-axis the stress expressed in B.U. A typical farinogram is shown in Figure 4.3.3. Specific parts of the graph are referred with different names, in particular one of them, the farinograph stability (the Stab parameter in the ISQ method) is measured in minutes and represents the time in which the higher line of the plot remains over the 500 B.U.

More in detail, the farinographic curve rises until it reaches a maximum peak that corresponds to the point at which the dough reaches its maximum consistency. From the moment the dough starts to form, it causes an ever increasing resistance to the rotation of the blades so the graph recorded by the pen point progressively widens given that it will have an oscillation that is proportional to the resistance. The maximum amplitude of the graph corresponds to the maximum resistance with which the dough opposes the rotation of the blades. As time proceeds, the resistance starts to decrease with a progression that the pen point will faithfully record. The farinographic stability expresses the resistance of the dough to the mechanical stress. High values of stability, in fact, correspond to doughs with high resistance to manufacturing.

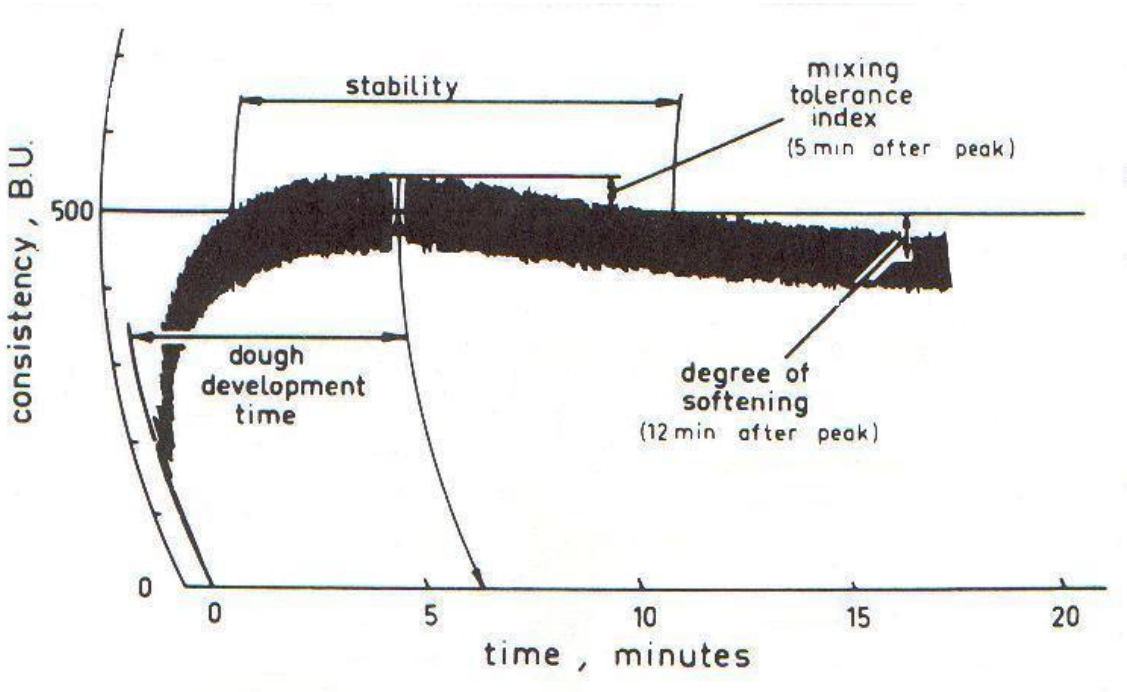


Figure 4.3.3. A typical farinographic curve.

The Falling Number (FN) values are obtained by means of an instrument (Figure 4.3.1c), based on principles of viscosimetry. The instrumental device used to measure the FN,

essentially determines the amylolitic activity of the wheat flour. Indeed the alpha-amylase enzyme works well as a catalyst, in fact, it attacks the complex glucose chains that make up the starch molecules and break them up, converting starch into reducing sugars and maltose. At room temperature and in presence of intact starch, the alpha-amylase enzyme activity is quite scarce. But the increasing temperature and the presence of damaged starch promotes an increase of the enzyme activity. The amylase enzyme always have a negative effect on the dough made from flour. Since alpha-amylase starts to be active in temperature range 55-80°C and affects damaged starches above all, the falling number determination uses the gelitinization (cooking) of the starches present in a suspension of water and flour. When the starches have become gelitinized, they are hydrolyzed by the enzyme, with a consequent liquefaction of the starch-water present in the suspension. The alpha-amylase activity is measured on this liquefaction.

Normally the measured final value of falling number (FN) of wheat doughs oscillates on average between 100-150 and 300 or higher. In other words, a FN value near, equal to, or greater than 300 indicates a weak or very weak alpha-amylase activity, in the interval between 200 and 250 indicates normal activity, while a FN values around 150-200, or less than 150 indicates that the alpha-amylase activity is high or very high.

Besides the FN, even the hectolitre weight (Phl) determination presents a minimum value for enter the ISQ classification. This parameter is determined by means of a Schopper balance, equipped with a ¼ liter vessel, and it furnishes information about the average density and the regularity of shape of the caryopses.

The SDS sedimentation index (also called SDS index) is expressed as the volume of gluten proteins (ml) precipitated in defined volume of 1.5% sodium dodecyl sulphate solution, according to the official method (A.A.C.C., 2000; Indrani et al., 2007; Delwiche et al., 1998).

It should be underlined that the most significant parameters by the point of view of the end-user, besides the protein content, determined by means of the traditional Kjeldhal method, are the alveographic indexes and the farinograph stability.

NIR analysis

Aliquots of each sample in the three physical forms were sent to 3 different laboratories for the acquisition of replicate NIR spectral measurements. The first laboratory (Lab. 1) is located at the Department of Life Sciences of the University of Modena and Reggio Emilia and it is equipped with a FT-NIR Bruker Optics MPA Multi Purpose Analyzer spectrophotometer. The second laboratory (Lab. 2) is located at CRA in S. Angelo Lodigiano

and it is equipped with a Foss NIRSystem 6500 spectrophotometer. The third laboratory (Lab. 3) is located at the Granaria Association in Milan and it is equipped with a Buchi NIRFlex N-500 spectrophotometer. In all the labs the integrating sphere sampling tool was used for spectra acquisition.

In Table 4.3.2. a summary of the characteristics of the spectral datasets acquired is reported.

Lab	**Sample form**	**repl #**	**sptr #**	**Range (cm^{-1})**
Lab. 1	F	4	1212	3850-12500
	S	4	1212	3850-12500
	G	4	1212	3850-12500
Lab. 2	F	2	604	4000-25000
	S	2	604	4000-25000
	G	2	606	4000-25000
Lab. 3	F	2-4	794	4000-10000
	S	2-4	824	4000-10000
	G	2-4	699	4000-10000

Table 4.3.2. Summary of the characteristics of the spectral datasets acquired in the different labs.

The use of instruments having different characteristics permitted to explore various wavenumber ranges (the FOSS spectrophotometer even reaches the visible region) and to use diverse resolutions, in a manner to appreciate also what were the ideal characteristics of a NIR instrument used for our aims. Figure 4.3.4 shows, for example purposes, the spectral data acquired on flour samples, in reflectance mode, by the three different spectrophotometer models.

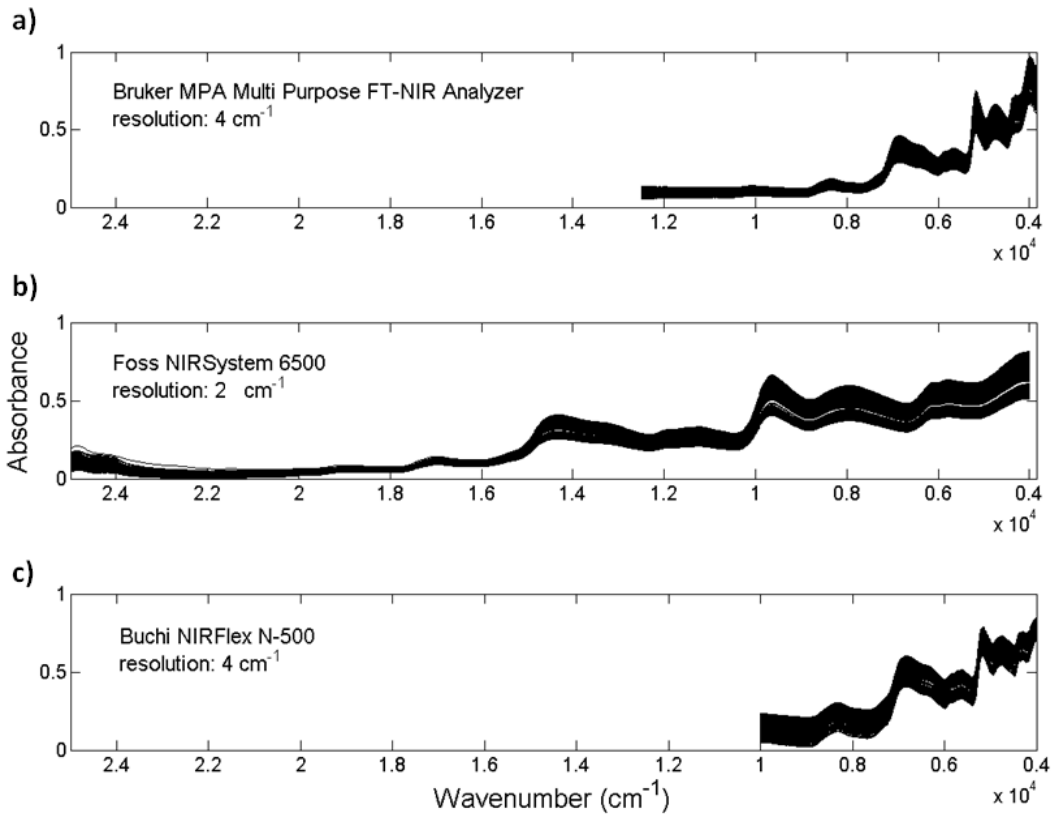


Figure 4.3.4. Spectra acquired in reflectance on white flour samples by the three spectrophotometer models used in Lab. 1 (a), Lab. 2 (b). and Lab. 3 (c).

Data Analysis

Statistical survey on chemical and rheological data

As a first step, a statistical survey on the seven chemical and rheological experimental parameters has been performed considering only the wheat samples (without considering the sample physical form). Mean value, range, standard deviation and experimental Root Mean Square Error (RMSE *Exp*) have been calculated for all the samples.

The reproducibility of the analyses carried out on each parameter was evaluated by calculating the RMSE *Exp* between replicates measurements, using the following equation 4.1:

$$RMSE_{Exp} = \sqrt{\left[\Sigma\left(\frac{V_{replicated\ values}}{N}\right)\right]} \qquad (4.1)$$

where V is the variance of the replicate measurements for each sample and N the number of replicated sets.

The RMSE Exp represents an index of the degree of variability inside the group of replicated analytical determination (laboratory experiments/measurements).

Explorative analysis on spectral data

The spectra acquired in each laboratory, for each sample form, were organized in nine independent datasets. The NIR datasets were, first of all, visually inspected in a manner to identify and eliminate the more noisy spectral regions, generally at the extremes of the spectral range.

Then, each meancentered dataset was submitted to explorative analysis using PCA. The samples lying outside the 99.7% confidence limit in the Q-T^2 plot were identified as outlier and removed from the datasets.

Finally, the spectra matrices were randomly divided in a training set (TRN) and in a test set (TST), paying care to maintain the replicate measurements of each sample in the same set: in this way the training set included the spectra of 203 samples and the test set the spectra of 100 samples. Considering the ISQ classes, the samples resulted to be divided as follows: 17 FAU, 20 FB, 37 FF, 81, FP and 48 FPS in the training set, and 7 FAU, 10 FB, 18 FF, 48, FP and 17 FPS in the test set.

Development of spectra based calibration models

Since we had to examine a number of datasets, the calibration model were calculated, considering different spectra pretreatments, being usually not possible to establish *a priori* which pretreatment works better (Xu et al., 2008). In particular, we considered the following pretreatments:

- none (N), meancentering (m), first derivative (d1), second derivative (d2), linear detrend (det1), Standard Normal Variate (SNV), that were tested both separately and in the following combinations:
- d1 + m, d2 + m, det1 + m, SNV + m.

An internal customized cross-validation was used for choosing the PLS models dimensionality. In particular, in the present application, a 4 cancellation groups cross-validation was used, forcing the algorithm to maintain the replicated spectra in the same group.

The performances in calibration (calculated on the training set, in cross-validation, and on the test set) were expressed in terms of correlation coefficients (R^2 cal, R^2 CV and R^2

Pred) and root mean squared errors (RMSEC, RMSECV, RMSEP). For any calculated model it was decided to not exceed 18 latent variables as maximum model dimensionality. The best model was chosen considering the lowest values of RMSECV preferring a model dimensionality as low as possible.

4.4 Results and discussion

Explorative data analysis

Statistical survey on chemical and rheological parameters

In Table 4.4.1 the statistical survey on the chemical and rheological parameters is presented. Concerning the farinograph stability (Stab), no value of RMSE Exp is presented because each wheat sample was analyzed only once. For each parameter, the RMSE Exp value resulted always much lower than the standard deviation value, proving the good reproducibility of the analyses. The RMSE Exp values were also used to obtain a direct comparison between the laboratory analyses and the results of the NIR based calibration models.

Parameter	**Phl (Kg/hL)**	**FN (sec)**	**Prot_ss% (%)**	**W (cm^2)**	**P/L**	**Stab (min)**	**SDS (mL)**
Mean value	78.59	345.79	13.56	244.18	0.62	11.27	73.39
Standard Deviation	4.10	66.66	1.30	102.74	0.84	6.54	9.79
Range	85.60-68.60	646.00-63.00	16.50-10.95	628.00-27.40	1.91-0.04	19.30-1.20	93.00-25.00
RMSE Exp	0.40	13.08	0.20	11.66	0.06	-	2.25

Table 4.4.1. Statistical survey on the experimental parameters.

Explorative PCA on chemical and rheological parameters

In order to identify the presence of outlier samples and to obtain preliminary information about sample clustering into the ISQ quality classes, PCA has been applied on the autoscaled chemical and rheological parameters dataset. Figure 4.4.1 reports the biplot of the first two PCs, that explain the 57.65% of variance. As it can be seen, the considered ISQ variables are quite explanatory of the quality differences between wheat categories, in fact the samples belonging to the different classes tend to cluster along the direction mostly defined by a strong contribution of PC1 (with a secondary contribution of PC2), even if there is a certain overlapping of the classes.

From the position of the variables in the biplot, it can be noticed that all the variables are influential on the definition the flour classes, since they are located at positive values of PC1 (Prot_ss%, SDS, Stab and W), as the FF samples and FPS samples.

Moreover some other considerations on Figure 4.4.1 could be made in order to explain the overlapping of the different end-use categories in the PC space. Indeed based of the ISQ parameters W, the FF samples can be distinguished from FPS samples even if the related clusters resulted overlapped. Notwithstanding the strong superimposition between samples belonging to categories FPS and FF, the W parameter has a great importance for the ISQ classification.

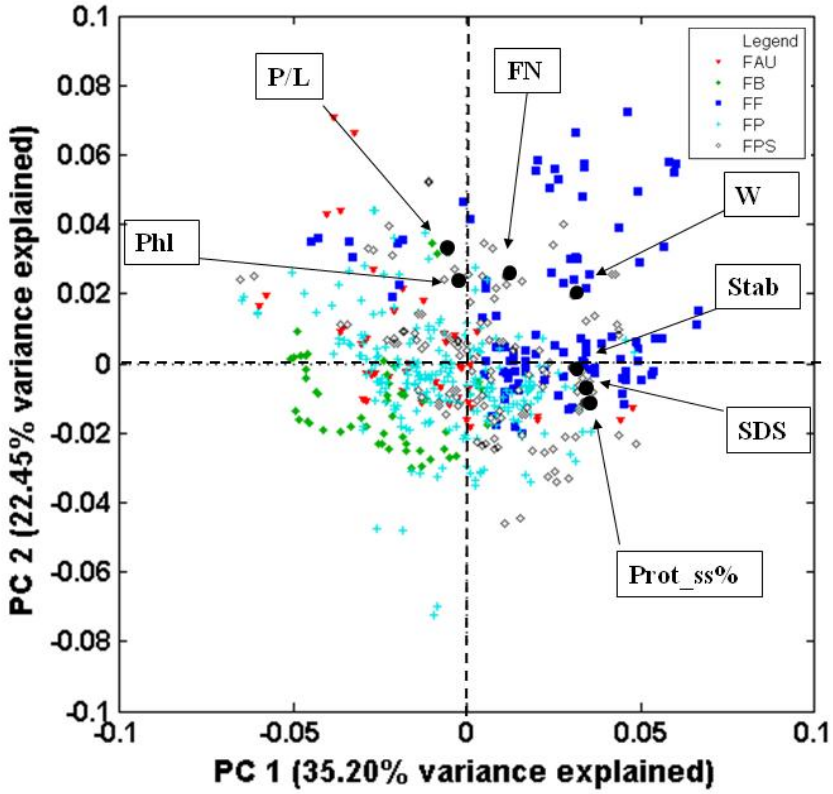


Figure 4.4.1. Biplot of the first two PCs for the ISQ parameters dataset.

Explorative PCA on NIR spectra datasets

Since we had to examine a number of datasets, individual explorative PCA models were calculated on meancentered NIR spectra measured by the different laboratories on the three cereal forms.

Table 4.4.2 reports the spectral ranges kept after visual inspection of the signals, the number of samples originally included in the TR and in the TS set, and the number of spectra deleted as outliers in both the sets after data exploration by PCA.

Lab	Sample form	Range kept (cm^{-1})	# TR sptrs	# TS sptrs	Outliers sptr
Lab. 1	F	3850-10500	812	400	8 TR/12 TS
	S	4000-10500	812	400	3 TR/14 TS
	G	3850-10500	812	400	19 TR/17 TS
Lab. 2	F	4000-25000	406	198	23 TR/12 TS
	S	4000-25000	406	200	19 TR/18 TS
	G	4000-25000	404	200	0 TR/3 TS
Lab. 3	F	4000-10000	530	264	21 TR/16 TS
	S	4000-10000	549	275	15 TR/1 TS
	G	4000-10000	474	225	17 TR/10 TS

Table 4.4.2. Summary of the spectral range kept for each dataset and of the subdivision of the samples in training and test sets.

To give an example, Figure 4.4.2 reports the Q-T^2 plot obtained for the PCA model concerning grit analysed by Lab. 2. As widely explained in Chapter 3, this kind of plot allows to evaluate the presence of outliers. In details, one spectrum presents anomalous T^2 values (higher than the 99.7% confidence limit), while twenty-four spectra presents T^2 values greater than the 95% confidence limit. Even if all these samples should be considered anomalous for their position inside the PCA model, only the sample outside the 99.7% confidence limit has been defined as outlier and removed. Concerning the Q residuals, five spectra are outside the 99.7% confidence limit, while twenty-seven spectra are outside of the 95% confidence limit. These spectra outside 99.7%, confidence limit are good candidates for outlier elimination but, with the purpose to limit as much as possible the sample removal, only the two spectra with higher values of Q residuals, indicated by the arrows in Figure 4.4.2, were removed.

Since all the PCA models obtained on the different datasets have shown similar results, only a single scores plot has been reported as an example. Figure 4.4.3 shows the PC1 *vs* PC2 scores plot for the wholemeal dataset acquired by Lab. 2. As it can be noticed, the first two PCs explained the 98.3% of the variance of the NIR dataset.

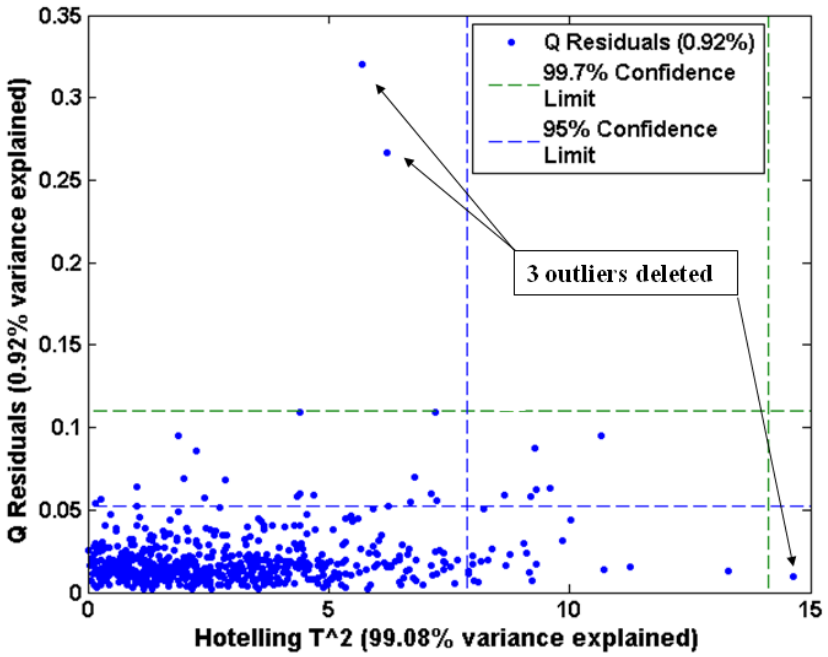


Figure 4.4.2. PCA model: Hotelling T2 versus Q residuals plot of Lab. 2. NIR spectral dataset related to wheat samples in grit form.

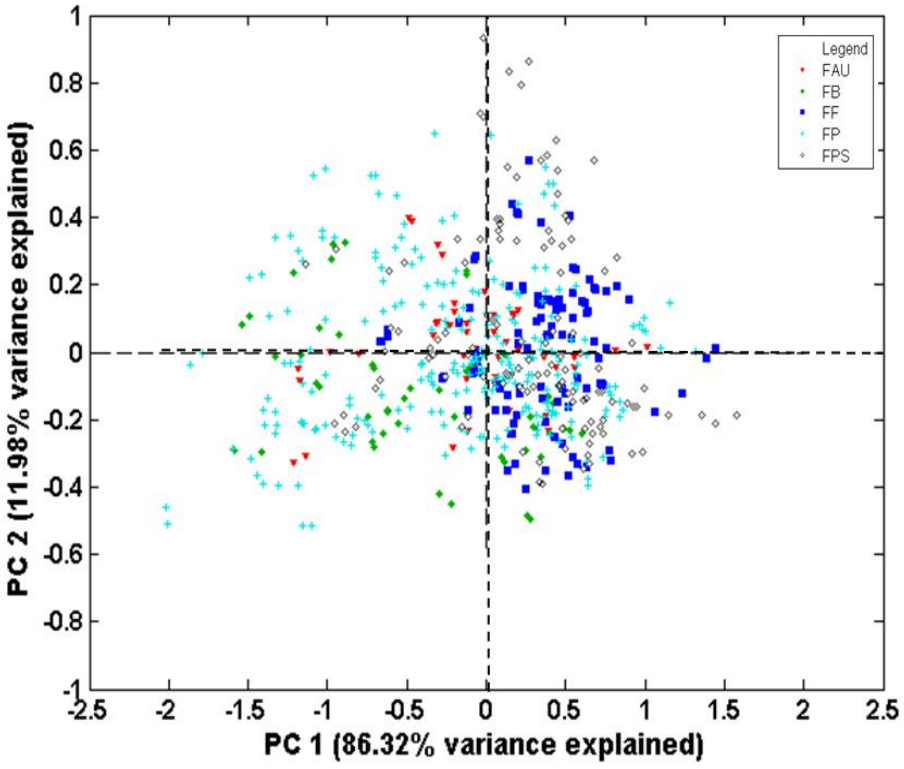


Figure 4.4.3. Scores plot of the first two PCs calculated by PCA on the mean NIR spectra of Lab. 2. NIR spectra related to samples in wholemeal form.

The NIR spectra seem to contain the chemical information able to classify the flour samples into different quality categories. Indeed the NIR spectra contain similar information with respect to that is brought by the ISQ parameters. In particular, the FP samples are spread in the whole PCs space and make the differentiation of the ISQ classes very difficult. However the FB and FF classes, which are the two mainly different ones as for their chemical and rheological properties, are visually distinguished along PC1 (FB samples are grouped at negative values of PC1, while FF samples at positive values of PC1).

As can be qualitatively seen by both the scores plots in Figures 4.4.1 and 4.4.3, the distinction of wheat samples into qualitative classes is not a trivial task (Foca et al., 2007; Foca et al., 2009; Cocchi et al., 2005).

PLS1 results

Tables 4.4.3, 4.4.4 and 4.4.5 report the results of the calibration models calculated by PLS1 for, respectively, flour, wholemeal and grit samples. For conciseness reasons, in the tables only the best models, obtained using selected pretreatments, have been included.

It has to be noticed that the dataset of wholemeal flour spectra acquired by Lab. 3 showed a clear differentiation of the samples based on the year of harvesting. So, before the building of PLS models, this dataset was submitted to the Orthogonal Signal Correction (OSC) (Wold et al., 1998), to remove from the matrix of signals such systematic variation, that is orthogonal to the information of interest.

In general, SNV and detrend, both singularly and in combination with meancentering, seemed to be the most promising pretreatments for the prediction of the ISQ parameters. As well known in literature (Shenk et al., 2001; Pojić et al., 2012), NIR spectroscopy, together with chemometric tools, is used for protein amount determination on cereal products. Indeed the PLS1 models related to protein percentage determination always showed high prediction performances also using different pretreatments on spectral data. The correlation coefficients in prediction resulted equal to 0.83 for white flour matrix, and it reached the maximum value equal to 0.95 for wholemeal flour matrix using the Lab. 2 spectral data. Relevant prediction results are obtained also for the hectoliter weight. In particular, the Phl model reached values of R^2 in prediction equal to 0.82 for the grit matrix using the Lab. 1 spectral data. The other PLS models on white and wholemeal flours showed R^2 Pred higher than 0.66 in both cases for Phl determination.

The PLS1 calibration models related to the rheological parameters, i.e., the falling number, the alveograph parameters W and P/L, the farinograph stability and the

sedimentation value, showed poor prediction performances independently of the physical form of the samples. Concerning white flour and wholemeal flour the PLS1 models related to the rheological variable W showed R^2 Pred equal to 0.59.

Lab	*Lab. 2*	*Lab. 3*	*Lab. 3*	*Lab. 1*	*Lab. 3*	*Lab. 3*	*Lab. 3*
ISQ variable	**Phl**	**FN**	**Prot_ss%**	**W**	**P/L**	**Stab**	**SDS**
Pretreatment	*SNV*	*SNV+ m*	*m*	*det1*	*d1+ m*	*det1+ m*	*det1+ m*
# of LVs	*16*	12	10	*16*	17	9	7
RMSEC	1.76	41.72	0.37	48.11	0.37	4.48	6.44
RMSECV	2.13	54.44	0.43	63.60	0.50	5.07	6.91
RMSEP	2.17	64.89	0.51	67.59	0.69	5.87	6.30
R^2 Cal	0.82	0.45	0.87	0.77	0.63	0.49	0.43
R^2 CV	0.74	0.14	0.83	0.61	0.37	0.36	0.35
R^2 Pred	0.68	0.05	0.83	0.58	0.37	0.25	0.33

Table 4.4.3. Best PLS1 models corresponding to the Min RMSECV for wheat as flour.

Lab	*Lab. 2*	*Lab. 2*	*Lab. 2*	*Lab. 2*	*Lab. 2*	*Lab. 2*	*Lab. 3*
ISQ variable	**Phl**	**FN**	**Prot_ss%**	**W**	**P/L**	**Stab**	**SDS**
Pretreatment	*det1+ m*	*SNV*	*d2+ m*	*det1*	*m*	*m*	*(OSC) det1*
# of LVs	*18*	*13*	*13*	*14*	*17*	*14*	6
RMSEC	1.35	47.56	0.15	52.94	0.43	4.57	6.44
RMSECV	1.73	53.50	0.19	63.17	0.52	5.30	6.63
RMSEP	2.34	54.39	0.31	66.02	0.69	5.54	7.02
R^2 Cal	0.90	0.43	0.98	0.70	0.63	0.51	0.43
R^2 CV	0.83	0.29	0.97	0.58	0.48	0.36	0.40
R^2 Pred	0.66	0.42	0.95	0.59	0.35	0.35	0.30

Table 4.4.4. Best PLS1 models corresponding to the Min RMSECV for wheat as wholemeal.

Lab	*Lab. 1*	*Lab. 3*	*Lab. 2*	*Lab. 3*	*Lab. 3*	*Lab. 1*	*Lab. 2*
ISQ variable	**Phl**	**FN**	**Prot_ss%**	**W**	**P/L**	**Stab**	**SDS**
Pretreatment	*SNV*	*det1+ m*	*d1+ m*	*SNV*	*SNV+ m*	*det1*	*det1+ m*
# of LVs	*15*	10	*16*	9	12	*8*	*14*
RMSEC	1.52	38.85	0.32	62.57	0.37	4.94	6.08
RMSECV	1.70	49.36	0.39	69.74	0.49	5.19	6.92
RMSEP	1.65	55.96	0.49	71.04	0.73	5.41	7.65
R^2 Cal	0.87	0.40	0.93	0.44	0.63	0.43	0.60
R^2 CV	0.84	0.11	0.90	0.32	0.39	0.37	0.49
R^2 Pred	0.82	0.04	0.87	0.30	0.30	0.34	0.32

Table 4.4.5. Best PLS1 models corresponding to the Min RMSECV for wheat as grit.

In order to obtain indications on the analytical performances of the labs, a comparison of the PLS1 models obtained for each variable was conducted, without considering the sample

physical form. The Table 4.4.6 reports the best PLS1 results for each parameter based on the absolute lower value of RMSECV. As it can be seen, the considered performances are quite explanatory of the quality differences between laboratories, in fact the spectra acquired by Lab. 3 gave better models for the determination of 4 parameters (FN, P/L, Stab and SDS) out of 7. On the other hand the Lab. 2 gained better results for the protein content and the rheological parameter W. Looking at the sample physical form, it has to be noticed that the spectra measured on grit and wholemeal flour samples furnished better results compared to the spectra acquired on white flour samples.

To give some examples, three measured-predicted plots of the models in Table 4.4.6 are shown in Figure 4.4.4. In particular, the plots of the models obtained for the protein content (Fig. 4.4.4a), the alveographic W (Fig. 4.4.4b) and the hectolitre weight (Fig. 4.4.4c) have been reported, representing with different colours the training set and the test set samples.

Lab/sample form	*Lab. 1/ G*	*Lab. 3/ G*	*Lab. 2/ S*	*Lab. 2/ S*	*Lab. 3/ G*	*Lab. 3/ F*	*Lab. 3/ S*
ISQ variable	**Phl**	**FN**	**Prot_ss%**	**W**	**P/L**	**Stab**	**SDS**
Pretreatment	*SNV*	*det1+ m*	*d2+ m*	*det1*	*SNV+ m*	*det1+ m*	*(OSC) det1*
# of LVs	*15*	*10*	*13*	*14*	*12*	*9*	*6*
RMSEC	1.52	38.85	0.15	52.94	0.37	4.48	6.44
RMSECV	1.70	49.36	0.19	63.17	0.49	5.07	6.63
RMSEP	1.65	55.96	0.31	66.02	0.73	5.87	7.02
R^2 Cal	0.87	0.40	0.98	0.70	0.63	0.49	0.43
R^2 CV	0.84	0.11	0.97	0.58	0.39	0.36	0.40
R^2 Pred	0.82	0.04	0.95	0.59	0.30	0.25	0.30

Table 4.4.6. Best PLS1 models corresponding to the Min RMSECV related to all different physical form samples. Comparison between laboratory performances.

Figure 4.4.4 confirms that all the samples are satisfactorily modeled as for protein amount, while the models concerning the hectolitre weight and, even more, the alveographic W resulted less satisfactory.

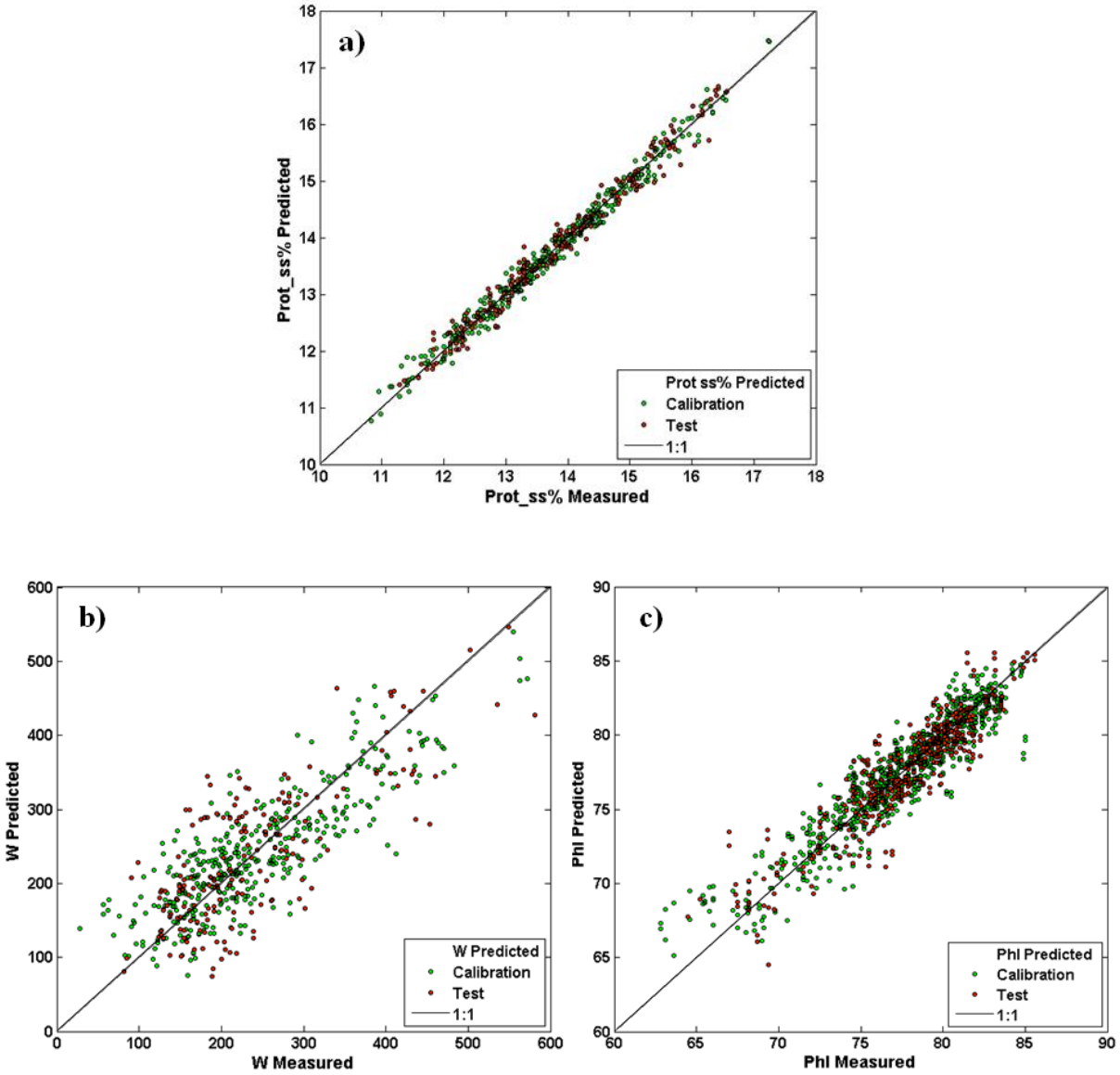


Figure 4.4.4. Measured-predicted plots for the best PLS1 calibration models obtained on proteins (a), alveographic W (b) and hectolitre weight (c).

PLS2 results

The best results of the PLS2 calibration models calculated on white flour, wholemeal flour and grit matrices are reported in Tables 4.4.7., 4.4.8. and 4.4.9., respectively. Among the several pretreatments tested, the best performing ones result to be similar in both PLS1 and PLS2 models. According to the PLS1 results, also in this case the protein content (prot_ss%) and the rheological parameter W showed high correlation coefficients in cross-validation.

However, it can be noticed that, on the whole, worst results both in calibration and in prediction were obtained by means of PLS2 compared to PLS1. PLS2 algorithm produces "overall" models able to predict all the chemical and rheological properties at the same time, requiring lower computational time. This is not surprising considering that PLS2 is based on

the calculation of a unique model for all the response variables selected and it is therefore selected as the best trade-off and not as the optimal model for the prediction of a particular property, e.g. the model dimensionality is defined as a compromise between the RMSECV values of all the variables. When variables having low correlation are considered, as in the case of the present study, this drawback becomes even evident.

Lab/Sample form	*Lab.2/ F*	*Lab.2/ F*	*Lab.2/ F*	*Lab.2/ F*	*Lab.2/ F*	*Lab.2/ F*	*Lab.2/ F*
ISQ variable	**Phl**	**FN**	**Prot_ss%**	**W**	**P/L**	**Stab**	**SDS**
Pretreatment	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*
# of LVs	*17*	*17*	*17*	*17*	*17*	*17*	*17*
RMSEC	1.96	50.06	0.43	55.32	0.51	4.73	6.46
RMSECV	2.34	58.32	0.47	65.10	0.57	5.23	7.65
RMSEP	2.21	63.63	0.51	61.98	0.65	5.25	7.21
R^2 Cal	0.78	0.43	0.88	0.71	0.59	0.48	0.55
R^2 CV	0.69	0.25	0.86	0.61	0.49	0.38	0.39
R^2 Pred	0.67	0.20	0.86	0.66	0.41	0.40	0.41

Table 4.4.7. PLS2 best model related to white flour matrix.

Lab/Sample form	*Lab.2/ S*	*Lab.2/ S*	*Lab.2/ S*	*Lab.2/ S*	*Lab.2/ S*	*Lab.2/ S*	*Lab.2/ S*
ISQ variable	**Phl**	**FN**	**Prot_ss%**	**W**	**P/L**	**Stab**	**SDS**
Pretreatment	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*	*det1+ m*
# of LVs	*18*	*18*	*18*	*18*	*18*	*18*	*18*
RMSEC	1.65	47.69	0.25	54.57	0.46	4.63	6.21
RMSECV	1.93	54.44	0.29	63.52	0.55	5.26	7.20
RMSEP	2.18	54.37	0.40	66.13	0.68	5.53	7.39
R^2 Cal	0.84	0.43	0.96	0.68	0.57	0.50	0.57
R^2 CV	0.79	0.28	0.94	0.58	0.42	0.37	0.43
R^2 Pred	0.70	0.42	0.91	0.59	0.36	0.35	0.35

Table 4.4.8. PLS2 best model related to wholemeal flour matrix.

Lab/Sample form	*Lab.1/ G*	*Lab.1/ G*	*Lab.1/ G*	*Lab.1/ G*	*Lab.1/ G*	*Lab.1/ G*	*Lab.1/ G*
ISQ variable	**Phl**	**FN**	**Prot_ss%**	**W**	**P/L**	**Stab**	**SDS**
Pretreatment	*SNV+ m*	*SNV+ m*	*SNV+ m*	*SNV+ m*	*SNV+ m*	*SNV+ m*	*SNV+ m*
# of LVs	*15*	*15*	*15*	*15*	*15*	*15*	*15*
RMSEC	1.71	54.11	0.40	67.61	0.50	4.87	6.86
RMSECV	1.87	58.12	0.44	74.42	0.57	5.27	7.54
RMSEP	1.81	56.90	0.50	74.71	0.68	5.45	7.85
R^2 Cal	0.83	0.30	0.89	0.55	0.49	0.44	0.45
R^2 CV	0.80	0.20	0.86	0.45	0.35	0.35	0.35
R^2 Pred	0.78	0.30	0.87	0.48	0.34	0.34	0.29

Table 4.4.9. PLS2 best model related to grit matrix.

4.5 Remarks

The results obtained in this work contain many useful suggestions for future development of the research in this field, even if the extraction of information on the technological quality of wheat by NIR signals is not a easy task. With respect to the traditional analyses, the spectra acquisition and spectra based calibrations are surely much more rapid, easier to use and could become widely diffused in very short times, since NIR spectroscopy is already extensively adopted for the quality control of many parameters in the cereals production/trade context.

The obtained results demonstrate that NIR spectroscopy could be a promising technique to these aims, in fact, the NIR transmittance spectroscopy has recently been employed to obtain the values for different alveographic and farinographic parameters, by means of specific calibrations (Miralbes, 2004; Corbellini et al., 2002). This fact demonstrates that the NIR spectrum contains the same information that were traditionally extracted from the "wet" analyses, suggesting that the chemical and rheological properties of bread wheat flour products can be partially evaluated from the corresponding NIR spectra. In particular, the protein content, among the parameters involved in the ISQ classification procedure, is a quantity rather simple to predict by means of NIR spectroscopy (Delwiche et al., 1998; Osborne, 1984; Osborne, 2008), as also evidenced by the numerous instruments on the market able to determine the protein percentage in cereals and cereal derivates. However, as for the rheological properties, the calibration curves obtained until now on flour or wheat grains are not so satisfactory and can be considered useful only for screening purposes (Delwiche and Weaver, 1994; Delwiche, 1998; Hruškova and Šmejda, 2003; Jirsa et al., 2008; Dowell et al., 2006).

In this thesis, satisfactory results for the prediction of the hectolitre weigh and the alveographic W have been obtained. In all cases PLS1 calibration models have shown better predictive performance compared to PLS2 models, both for the R^2 and the RMSE values obtained. Concerning a comparison of the three laboratory, the chemometric analysis evidenced a higher performance of Lab. 3 compared to the others, for which generally better PLS1 models have been obtained.

4.6 Acknowledgements

The Consiglio per la Ricerca e la sperimentazione in Agricoltura (CRA) in S. Angelo Lodigiano and the Granaria association in Milan are acknowledged for the execution of part of the NIR analyses presented in this work.

4.7 References

A.A.C.C. (2000). SDS index: official method. Approved methods (10th ed.). American Association of Cereal Chemists. St. Paul, MN.

Alava, J.M., Millar, S.J., Salmon, S.E., (2001). The Determination of Wheat Bread making Performance and Bread Dough Mixing Time by NIR Spectroscopy for High Speed Mixers. *J. Cereal Sci,* 33, 71-81.

ASS.IN.CER. (Associazione Intersettoriale Cereali e altri Seminativi) birth in Bologna in 1996. www.assincer.it.

Barton, II. F.E., Shenk, J.S., Westerhaus, M.O., Funk, D.B., (2000). The development of near infrared wheat quality by locally weighted regressions. *J. NIRS* 8, 201-208.

Borasio E., (1997). Classificazione merceologica del frumento con indici di qualità. *Agricoltura.* 9, 59-61.

Borneo, R., Khan, K., (1999). Protein changes during various stages of bread making of four spring wheats: quantification by size exclusion HPLC. *Cereal Chem.*, 76 (5), 711-717.

Cappelli, P., Vannucchi, V., (2000). *Cereali e derivati*. Chimica degli alimenti. Ed. Zanichelli.

Corbellini, M., Empilli, S., Lucisano, M., Pagani, M.A., (2002). Metodi rapidi di analisi del grano tenero. *L'Informatore Agrario*. LVIII. (33), 53-54.

Cocchi, M., Corbellini, M., Foca, G., Lucisano, M., Pagani, M.A., Tassi, L., Ulrici, A., (2005). Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra. *Anal. Chim. Acta*. 544, 100-107.

Coultate, T.P., (2005). La chimica degli alimenti. Ed. Zanichelli, 1st Italian Ed. based on the 3rd English Ed. (Royal Society of Chemistry), 37-44.

D'Apollonia, B.L., Kunerth, W.H., (1984). The farinograph handbook. Am. Ass. Cereal Chem., St. Paul, MN.

Delwiche, S.R., (1998). Protein Content of Single Kernels of Wheat by Near-Infrared Reflectance Spectroscopy J. *Cereal Sci*, 27, 241-254.

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

Delwiche, S.R., Weaver, G., (1994). Bread quality of wheat flour by near-infrared spectrophotometry: feasibility and modeling. *J. Food Sci.* 59, 410-415.

Dowell, F.E., Maghirang, E.B., Xie, F., Lookhart, G.L., Pierce R.O., Seabourn B.W., Bean, S.R., Wilson, J.D., Chung, O.K., (2006). Predicting Wheat Quality Characteristics and Functionality Using Near-Infrared Spectroscopy. *Cereal Chem*., 83 (5), 529-536.

Danno, G., Honesey, R.C., (1982). Changes in flour protein during dough mixing. *Cereal Chem.* 59 (4), 249-253.

Fennema, O.R., (1996). Cereals and cereal products. Food Chemistry, Ed. Marcel Dekker, 631-654.

Foca, G., Ulrici, A., Corbellini, M., Pagani, M.A., Lucisano, M., Franchini, G.C., Tassi, L., (2007). Reproducibility of the Italian ISQ method for quality classification of bread wheats: An evaluation by expert assessors. *J. Sci. Food Agric*. 87 (5), 839-846.

Foca, G., Cocchi, M., Li Vigni, M., Caramanico, R., Corbellini, R., Ulrici, A., (2009). Different feature selection strategies in the wavelet domain applied to NIR-based quality classification models of bread wheat flours. *Chemom. Intell. Lab. Syst*. 99, 91-100.

Foca, G., Ulrici, A., Corbellini, M., Pagani, M.A., Lucisano, M., Franchini, G.C. Tassi, L., (2007). Reproducibility of the Italian ISQ method for quality classification of bread wheats: An evaluation by expert assessors. *J. Sci. Food Agric.* 87, 839-846.

Greer, E.R., Ziegler, E., (1974). Principles of milling. In: Wheat: Chemistry and Technology. Ed. by Pomeranz, American Association of Cereal Chemists, St. Paul, MN, 115-200.

Graveland, A., Bosveld, P., Lichtendonk, W.J., Moonen Hans H.E., and Scheepstra, A., (1982). Extraction and Fractionation of Wheat Flour Proteins. *J. Sci. Food Agric.* 33, 1117-1128.

Hermansson, A.M., Svegmark, K., (1996). Developments in the understanding of starch functionality. *Trends Food Sci. Tech.* 7, 345-353.

Hinton, J.J.C., (1959). The distribution of ash in wheat kernel. *Cereal Chem.* 36, 19-31.

Hruškova, M., Šmejda, P., (2003). Wheat flour dough alveograph characteristics predicted by NIRSystems 6500. *Czech J. Food Sci*, 21, 28-33.

Indrani, D., Sai Manohar, R., Rajiv J., Venkateswara R.G., (2007). Alveograph as a tool to assess the quality characteristics of wheat flour for parotta making. *J. Food Eng.* 78, 1202-1206.

Jirsa, O., Hruškova, M., Švec, I., (2008). Near-infrared prediction of milling and baking parameters of wheat varieties *J. Food Eng.* 87, 21-25.

Lehninger, A.L., (1975). Biochemistry. 2nd Ed. Worth Publisher Inc.

Miralbes, C., (2004). Quality control in the milling industries using near-infrared transmittance spectroscopy. *Food Chem.* 88 (4), 621-628.

Miles M.J., Morris, V.J., Orford, P.D., Ring, S.G., (1985). The roles of amylose and amylopectin in the gelation of starch. *Carbohydr. Res.* 135, 271-281.

Morris, V.H., Alexander, T.L., Pascoe, E.D., (1945). Studies of the composition of the wheat kernel I. Distribution of ash and protein in center sections. *Cereal Chem.* 22, 351-361.

Osborne, B.G., (2008). In: Burns D.A., Ciurczac E.W. (eds) Handbook of Near Infrared Analysis, 3rd edn. CRC Press, Boca Raton, FL.

Osborne, B.G., (1984). Investigations into use of near infrared reflectance spectroscopy for the quality assessment of wheat with respect to its potential for bread baking. *J. Sci. Food Agric.* 35, 106-110.

Pagani, M.A., Lucisano, M., Mariotti, M., (2002). Valutazione del grado di gelatinizzazione dell'amido mediante tecnica NIR. *Tecnica Molitoria*. 53, 1218-1223.

Palmer, G.H., (1989). Wheat in milling and baking. Cereal Science and Technology. Aberdeen University Press, Aberdeen, 378-379.

Pasqualone, A., Piergiovanni, A.R., Laghetti, G., Volpe. N., Simeone, R., (Ottobre 2006) Valutazione di pane ottenuto da grano kamut e da spelta. *Tecnica Molitoria*. 1075-1080.

Pojić, M., Mastilović, J., Majcen, N., (2012). The Application of Near Infrared Spectroscopy in Wheat Quality Control. Infrared Spectroscopy - Life and Biomedical Sciences. 11, 168-185 Prof. Theophanides Theophile (Ed.) InTech.

Posner, E.S., (2000). Wheat. Handbook of Cereal Science and Technology, 2nd Ed. by Kulp & Ponte, Marcel Dekker, 1-29.

Schofield, J.D., Bottomley, R.C., Timms, M.F., Booth, M.R., (1983). The effect of heat on wheat gluten and the involvement of sulphydryl-disulfide interchange reactions. *J. Cereal Sci.* 1 (4), 241-253.

Sgrulletta, D., De Stefanis, E., (1997). Simultaneous evaluation of quality parameters of durum wheat (*Triticum Durum*) by Near Infrared Reflectance Spectroscopy. *Ital., J. Food Sci.* 9, 295-301.

Singh, N., Singh, H., Singh B., (1998). Determining the distribution of ash in wheat using debranning and conductivity. *Food Chem.* 62 (2), 169-172.

Shuey, W.C., Tipples, K.H., (1982). The amylograph handbook. Am. Ass. Cereal Chem., St. Paul, MN.

Shenk, J.S., Workman, J.J.,. Westerhaus, M.O., (2001). in: Burns D.A.,. Ciurczak E.W., (Eds.), Application of NIR spectroscopy to agricultural products. In: Handbook of Near-Infrared Analysis, Marcel Dekker, New York, US, 419-474.

Wesley, I.J., Larroque, O., Osborne, B.G., Azudin, N., Allen, H., Skerritt, J.H., (2001) Measurement of gliadin and glutenin content of flour by NIR spectroscopy. *J. Cereal Sci.* 34, 125-133.

Wesley, I.J., Uthayakumaran, S., Anderssen, R.S., Cornish, G.B., Bekes, F., Osborne B.G., Skerritt, J. H., (1999). A curve-fitting approach to the near infrared reflectance measurement of wheat flour proteins which influence dough quality. *J. NIRS* 7, 229-240.

Wesley, I.J., Larsen, N., Osborne, B.G., Skerritt, J.H., (1998). Predicting Protein Composition, Biochemical Properties, and Dough-Handling Properties of Hard Red Winter Wheat Flour by Near-Infrared Reflectance. *J. Cereal Sci.* 27, 61-69.

Wold, S., Antti, H., Lindgren, F., Ohman, J., (1998). Orthogonal signal correction of near-infrared spectra. *Chemom. Intell. Lab. Syst.* 44, 175-185.

Xu, L., Zhou, Y.P., Tang, L.J., Wu, H.L., Jiang, J.H., Shen, G.L., Yu, R.Q., (2008). Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Anal. Chim. Acta*. 616, 138-143.

Chapter 5: SWINE ADIPOSE TISSUE: CLASSIFICATION OF SAMPLES FROM DIFFERENT SUBCUTANEOUS LAYERS AND NIR-BASED PREDICTION OF FAT COMPOSITION

5.1 Introduction

A very peculiar aspect of the swine fat covering tissue is that it is constituted by two layers having different composition and coming from different metabolic ways; these layers are generally easy to distinguish, since they present visibly different color and consistency, and may be physically separated by means of a simple cut. The outer layer, close to the rind, presents a greater consistency and is richer in unsaturated fatty acids with respect to the inner layer (Malmfors et al., 1978; Girard et al., 1988). The greater consistency of the outer layer, despite the higher degree of unsaturation, is probably due to a greater collagen content, i.e., the connective tissue, and to a greater organization of the connective structure surrounding the adipocytes (Lo Fiego et al., 1987).

The stratification of the pig fat in different layers plays an important role in the Italian industry, in fact the processing industry makes use of the outer fat, considered harder, in the form of cubes for the production of salami and sausages (Santoro, 1984), while the inner layer tissue, classified as soft fat, is mainly destined to the melting. Normally a trained and expert operator is able to separate the two subcutaneous layers by hand-slashing, since the mechanical on-line separation between layers results difficult and uncertain.

The organoleptic and technological quality of the fat, reflecting the consistency of the fat belonging to each layer, may be estimated by means of a number of chemical analyses, such as the iodine value analysis and the gas chromatographic determination of the fatty acids composition. These common methods are, however, expensive, time-consuming, detrimental for the environment, because of the chemical reagents involved, and they are not suitable to be used to follow an industrial process in real-time.

In this work, therefore, we applied fast and reliable analytical methods coupled with chemometric techniques of data analysis to address two different issues in the meat industry:

i. classification of fat samples coming from two different subcutaneous layers to be destined to different end-uses;

ii. fast determination of iodine value and fatty acids composition of fat samples.

Hence, after the paragraph 5.2, where some information on the characteristics of the products of swine origin are introduced, the present Chapter of this thesis is divided into two main parts (5.3 and 5.4), where the two separate problems here presented are tackled.

More in detail, in paragraph 5.3, the cheap and fast analytical techniques represented by tristimulus colorimetry and FT-NIR spectroscopy coupled to chemometrics have been used to classify 205 samples of fat, collected from 66 pig specimens at two different depths below the rind.

In the literature, colorimetric measurements have been used to investigate possible relationships between fat color and fat composition (Wood et al., 2003), that is often variable, depending on genotype and sex of the pig and on the rearing system adopted. (Carrapiso and Garcia, 2005) proved that the colorimetric characteristics of subcutaneous fat samples are closely related to their fatty acids composition; in particular, the largest correlationship involve L* which is negatively related to most unsaturated fatty acids and positively related to the most abundant saturated fatty acids. (Gandemer, 2002), indeed, found higher degrees of whiteness and pinkness in firm fat than in low consistency fat. As for the NIR-related literature, the papers by Pérez-Juan et al. (2010) and Zamora-Rojas et al. (2013) merit a special mention for our aims, since they analyzed by NIR, with satisfactory results, the fatty acids composition at two different locations of subcutaneous fat.

The second part of the work, reported in paragraph 5.4, concerns the evaluation of the potential of NIR spectroscopy, coupled to proper multivariate calibration methods, to predict the iodine value and the fatty acids composition of fat samples, starting from reference measurements acquired by the traditional Wijs method and gas chromatographic analysis.

A number of papers are present in literature about similar topics. According to (Li et al., 1999) and (Cox et al., 2000) the iodine value was predicted by FT-NIR spectroscopy on melted fat and oil samples with good reproducibility compared to the traditional Wijs method. Moreover, there are some interesting works related to fatty acids determination based on NIR measurements acquired by means of a fibre optic probe (Pérez-Marín et al., 2009; González-Martín et al., 2003), while other works have used diffuse reflectance, transmission and transflectance measurements (Gjerlaug-Enger et al., 2011; Müller and Scheeder, 2008; Ripoche and Guillard, 2001); all of them reported good results.

As for the chemometric techniques used, a preliminary data exploration has been always performed by means of PCA, that allowed to remove outlier data. To face the classification and calibration issues, PLS-DA and PLS have been initially used as classification and calibration methods, respectively; then, also variable selection algorithms have been applied

to the different datasets. As widely explained in Chapter 3, since it is not possible to know in advance with pretreatment is more effective to extract the useful information from spectra, several combination of pretreatments have been investigated in both the works to obtain the best models.

A thorough discussion on the obtained models have been finally performed, with different purposes. Among them:

- to individuate the operating conditions (both considering the instrumental conditions and the chemometric procedures adopted) which have led to obtaining the best models;
- to understand which are the most informative spectral regions for the different objectives of the work;
- to get a better understanding of the chemical characteristics of the different types of swine fat;
- to understand the reasons of the misclassification of certain samples in the classification part of the work;
- to understand which fatty acids may actually be quantified with NIR spectroscopy in a real process.

5.2 Swine meat and adipose tissue

The meat is defined as the part of the animal used as food. Normally is also the greater edible part (in weight) of animal body. The most important species for the production of meat remain domestic cattle, sheep, swine and poultry. Cattle, sheep and swine are often referred to as a species from the "red meat", while poultry as "white meat". The importance of the three species of red meat in the supply of meat proteins, differs in different parts of the world. For example, beef is more important in North and South America, Africa and Europe, while the sheep are the most important in the Near East Europe and swine in the Far East Europe and Italy (Warriss, 2000).

Consumption of meat and meat derivates in Italy

In the early 50's meat annual consumption in Italy had a positive increasing, and over the last forty years it has been observed a constant growth in consumption of all kinds of meat and other animal protein source. The early stages of the growth of meat annual consumption have been particularly striking, because it was based on the situation of absolute shortage after the second world war. The increased annual consumption of meat, as a whole, was

particularly evident from 1950 to 1970. For the swine meat, in particular, after a slow start (6.6 kg / person for the years 1955-1960), the increase was registered (8.9 kg over the years '61-'70, 16 kg in the years '71-80, 29 kg in 1999 up to 30.0 kg in 2002). The increase was primarily due to strong growth in consumption of salami and processed meat products, rather than to the fresh swine meat. However, this trend took advantages of the events and diseases involving other edible animal species, such as BSE for cattle production in ’80 and in 2000, bird diseases for poultry production in ’90 and more recently in 2003-2005 and 2009 in relationship with bird flues called AH5N1 and AH1N1 respectively (Rimoldi, 2011).

Year	Production (# 10^3 Tons)	Import (# 10^3 Tons)	Export (# 10^3 Tons)	Availability (kg / person)	Consumption (kg / person)
2012	1271	979	205	1848	30.7
2011	1277	1051	205	1874	31.1
2010	1336	1040	189	1923	31.9
2009	1306	917	173	1848	30.8
2008	1291	902	184	1843	30.8
2007	1274	1000	156	1867	31.3
2006	1234	967	149	1818	30.9
2005	1200	905	148	1776	30.2
2004	1203	926	152	1775	30.3
2003	1181	924	128	1771	30.3
2002	1158	892	120	1750	30.0

Table 5.2.1. Italian production, import, export, annual availability and annual consumption/person of swine meat and swine transformed products (salami and hams) expressed in thousands of tons and kg/person, respectively.

Table 5.2.1, based on ISTAT data and elaborated by ASS.I.CA (ASS.I.CA., 2007-2012), shows the annual consumption of swine meat in the last decade (2002-2012). It is evident that meat consumption is still fixed at high level constantly (greater than 30.3 kg / person). The same situation can be registered also in Europe. In fact, in almost all EU countries, the average consumption / person resulted higher.

This trend means also that the industries have evolved in the last decade and have gradually adapted to the demands of different consumers. Moreover also the breeding techniques gradually changed in a way to obtain defined starting materials (i.e., ‘heavy’ pigs as live animals based on their weight) for selected food productions, such as Protected Designation of Origin (PDO) foodstuffs.

The production of the Italian heavy swine aims essentially to provide thighs for the production of PDO dry-cured hams, such as Parma ham (PDO Recognition: Reg. CE n.1107/96. www.assica.it). PDO laws for certified food products are severe and Consortia for the protection of PDO product also dictated rules for the required characteristics for the fresh

thighs, the genotypes (breeds and crosses) that are allowed, the age and the slaughtering weight of pigs, and the feed that can be used.

The genotype rules admit only some pure bred subjects, or hybrids obtained from some breeds. As pure bred, only individuals from the Italian Large White and Italian Landrace breeds can be used. In addition, the crosses with the Italian Duroc breed are permitted (Bosi et al., 2004).

Some nutritional aspects of swine meat and fat

Nowadays the consumers are more and more interested in improving the quality of their life which is a function of proper nutrition and consumption of healthy foods with a high biological value (Wilfart et al., 2004). Moreover in 2002 also the international organization WHO (World Health Organization) concluded that diet, in addition to lifestyle, played an important role on the process of developing certain diseases, such as cardiovascular disease, stroke, obesity, diabetes mellitus, hypertension, and cancer (WHO/FAO., 2002). This hypothesis is in agreements to the consumers point of view.

In the developed countries the cardiovascular diseases represent the first cause of mortality. However, some of their main risk factors are closely related to the diet, to the amount of fatty acids present in the food, to the chemical composition of these lipids (saturated, monounsaturated, polyunsaturated) (Murray et al., 1997) and to the high plasma levels of total cholesterol and its main carrier, low density lipoproteins (LDL) (Cordain et al., 2005; Ashwell, 1993).

Currently, the swine meat is one of the food most consumed by the population in Italy. It represents a high value source of vitamins, proteins, fat and oligo-elements recommended in the diet of each human growing phase (from male and female childhood to grown adults), according to the Italian INRAN (www.inran.it) and SINU (www.sinu.it) institutions which are involved in the research about human nutrition. Swine meat has an important place in a healthy diet, providing proteins with a good balance of amino acids, rich in essential amino acids, iron in a ready available form, B group vitamins and other essential minerals. Meat also contributes a significant amount of triglycerides to the human diet and it is this component that has been most under the spotlight in recent years in relation of the healthiness of people consuming meat. Indeed, the heavy swine meat contains an amount of saturated fatty acids (SFA), mono unsaturated fatty acids (MUFA) and poly unsaturated fatty acids (PUFA) in a fraction equal to about 1/3 of the total lipids weighed, respectively (Table 5.2.2). The total amount of lipids in swine meat results lower in weight compared to the past, corresponding in

an improvement of meat quality in terms of human healthiness. However, the FA composition in meat can be greatly modified by production factor such as animal diet, feeding regimes, age, weight, sex, breed (Bosch et al., 2012; Lo Fiego et al., 2010; Webb et al., 2008; Bosi et al., 2004), species and site of the animal body (Lo Fiego et al., 2005). Moreover also the amount of vitamin A contained in swine adipose tissue can be modified by a proper addition to feeding regimes (Minelli et al., 2013).

From a biological point of view, the saturation of FA dictates the melting point of a adipose tissues (firmness): highly saturated triglycerides, mainly composed by SFA, have a higher melting point (firmer) than unsaturated triglycerides. Swine dietary fats and carbohydrates are the sources of long chain FA for synthesis of lipids in mammals (Mayes, 1996). Dietary fats are readily converted to subcutaneous and/or intramuscular fat tissues. The adipose tissues formed in this manner takes the general characteristics of the dietary fat. Dietary carbohydrates are converted to swine body fat through a process called "de novo fatty acid synthesis," forming predominantly SFA and MUFA (Mayes, 1996), which yield a firmer adipose tissues.

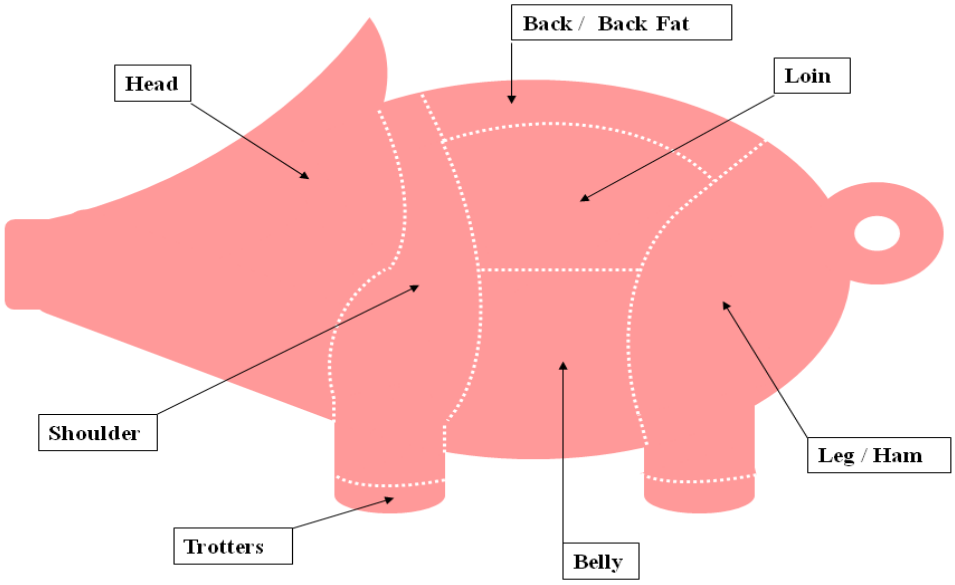
Swine organism requires the essential fatty acids (i.e., linoleic acid C18:2) from a fat source in the diet to incorporate PUFA into the adipose tissues of the carcass. Dietary fat additions will alter or even shut down de novo fat synthesis (Mayes, 1996). Thus, as the percentage of fat is increased in the diet, de novo fatty acid synthesis is further inhibited, resulting in less saturated fat deposition (softer). Furthermore, as the fatty acid profile of swine dietary fat becomes less saturated (softer), swine adipose tissues also becomes less saturated (softer).

Indeed, triglycerides rich in oleic (C18:1) and linoleic acids have lower melting points, since the two single fatty acids blend between 13 °C and -5 °C, while the corresponding saturated acids, such as palmitic (C16) and stearic (C18) acids, have melting points between 63 and 70° C (Russo et al., 1989). Moreover, even the change in the molecular structure of the individual fatty acids is important. In fact, *trans* fatty acids melt at higher temperatures than their *cis* isomers, as well as those with a linear chain with respect to acids with branched structure (Enser, 1984).

For these reasons, in recent decades, the swine production in Italy have changed in a way to obtain products with limited amounts of adipose tissues (and total cholesterol). This included also a variability in fat composition by replacing SFA with polyunsaturated fatty acids, producing an increase of linoleic acid (C18:2). (Khosla and Hayes, 1994) have shown that large amounts of linoleic acid (greater than 10%) are able to reduce not only LDL, but the

high density lipoproteins (HDL) with an improvements in human healthiness and longevity (Cordain et al., 2005).

The average FA composition and the corresponding percentage in different swine body tissues are shown in Table 5.2.2. obtained from the data elaborated by INRAN (Italian national institute for food and nutrition researches) and SINU (Italian society of human nutrition) (INRAN, 2003). The same different swine tissues (cuts) are represented in Scheme 5.2.1.



Scheme 5.2.1 Swine fresh meat: primal cuts.

	Heavy Pig. Italian Domestic swine: Fresh meat	Food Code **Leg**: 105600	Food code **Shoulder**: 105800	Food Code **Loin**:105700
	Chemical composition	**Value for 100g of edible part**	**Value for 100g of edible part**	**Value for 100g of edible part**
Chemical composition and energy values	**Edible Part (%):**	90	94	78
	Water (g):	72.9	70.6	68
	Proteins (g):	20.4	19	20.8
	Lipids (g):	5.1	8.9	9.9
	Colesterol (mg):	89	83	88
	Total Energy (kcal-KJ):	128-533	156-653	172-721
	Total Energy ratio	64% in Proteins + 36% in Lipids + 0% in Carbohydrates	49% in Proteins + 51% in Lipids + 0% in Carbohydrates	48% in Proteins + 52% in Lipids + 0% in Carbohydrates
	Sodium Na (mg):	76	73	59
	Potassium K (mg):	370	220	300
	Iron Fe (mg):	1.7	1.2	1.4
	Calcium Ca (mg):	8	6	7
	Phoshorus P (mg):	176	150	158
	Vitamin B1 Thiamine (mg):	0.31	0.26	0.28
	Vitamin B2 Riboflavin (mg):	0.31	0.34	0.3
	Vitamin PP Niacin (mg):	3.8	3	3.8
	Vitamin A Retinol eq. (µg):	5	6	6
	FA composition	**g/ 100 g of Leg edible part**	**g/ 100 g of Shoulder edible part**	**g/ 100 g of Loin edible part**
Fatty Acids Composition	**Total Lipids (%):**	5.1	8.9	9.9
	Saturated Fatty Acids SFA (%):	1.72	2.97	3.5
	C12:0	0.02	0.06	0.02
	C14:0	0.05	0.11	0.13
	C16:0	1.03	1.77	2.06
	C18:0	0.58	1	1.18
	Mono Unsaturated Fatty Acids MUFA (%)	1.99	3.54	3.87
	C16:1	0.09	0.16	0.16
	C18:1	1.87	3.29	3.57
	C20:1	0.02	0.09	0.05
	Poly Unsaturated Fatty Acids PUFA (%)	0.87	1.63	1.54
	C18:2	0.6	1.33	1.06
	C18:3	0.03	0.1	0.08
	C20:4	0.17	0.18	0.12
	C20:5	0.03	0.02	0.01
	PUFA / SFA ratio:	0.5	0.5	0.4

Table 5.2.2. Tables of swine fresh meat composition.

The role of fat in meat production and meat industry

Since the fat in swine meat is mainly located in a tissue visible to the naked eye, i.e., subcutaneous covering fat and/or intramuscular fat, it is easy to be physically separated from the protein part of the meat. This represents an advantage of the swine meat able to be used for low caloric diets and/or for selected high quality hand cured productions such as PDO Parma ham.

Nowadays consumers want to have food that aid in a healthy lifestyle. High quality food may represent a starting point for longer and better existence. In the common and widespread opinion, the healthy existence seems to be related to the consumption of reduced fat content foods (Wilfart et al., 2004; Wood et al., 2008). Consumer appreciation of fat in a food is a combination of different aspects: on the one hand, the consumers refuse any visible fat, on the other hand, they consider that the quality of cooked foods having an high fat content is better, because they present considerable improvements in flavor properties (Janz et al., 2009).

For the same reasons the presence of separated adipose tissues have a role in meat product quality, contributing to texture (tenderness and mouthfeel) and juiceness in both fresh meat and cooked meat products. The consistency of adipose tissue (represented by its firmness/softness/hardness) is influenced by FA composition. Such a composition affects some physical and rheological properties such as the sliceability of meat products or emulsion stability in meat transformations by heat. Fatty acid composition of subcutaneous adipose tissues and intramuscular fat contribute to the meat fragrance, flavor and odour. Indeed flavour and odour are closely associated. Flavour is mainly determined by water-soluble constituents, odour by fat-soluble, volatile elements. Fragrance and flavor of transformed meat products (cooked or seasoned) are strongly affected by oxidation of lipids. For example, during cooking some volatile products of lipid oxidation involved in the Maillard reaction are obtained. They represent additional volatile compounds that contribute to the flavor and smell of the product (Wood et al., 2003).

The chemical and physical properties of fat influence the eating and keeping qualities of meat (Wood et al., 2008; Bosi et al., 2004). Reducing the fat content in meat may adversely affect the eating satisfaction. Saturated fats containing long-chain fatty acids solidify easily upon cooling and will, therefore, affect the palatability of meat. Palatability (i.e., eating quality that encompasses three main characteristics: texture, juiciness and flavour/odour) is the primary determinant of consumer acceptance at the point of consumption. The degree of saturation of fat is one of the most important characteristics that influence the quality

parameters. Intramuscular fat and moisture affect juiciness (in good balance between dryness and succulence) and flavor directly and tenderness indirectly (Webb et al., 2008).

The fat tends to deteriorate by mean of two different processes: the hydrolysis of mono-, di- and tri-glycerides and the oxidation of fatty acids. The hydrolysis process leads to the liberation of glycerol and free fatty acids (long-chain carboxylic acids), weakly acidic, which catalyze the further reaction of esterification. Moreover, the hydrolysis increases the speed of the oxidative degradation. Indeed, free fatty acids contain a carboxyl group that can react with peroxides in a way to form free radicals, that are promoters of oxidation. The oxidation process is a transformation of fatty acids leading to the formation of volatile substances. In particular, it proceeds much more rapidly as greater is the degree of unsaturation in fatty acids.

The effect of fatty acids on shelf life and meat quality aspects (Wood et al., 2008; Bosi et al., 2004) are explained by the propensity of unsaturated fatty acids to oxidize, leading to the development of rancidity as display time increases. The chemical oxidation of unsaturated fatty acids by means of gaseous oxygen releases peroxides with free radicals which may react to damage proteins, enzymes, other lipids and vitamins. The FA oxidation with O_2 is a autocatalytic reaction (radical mechanism) and once started the reaction rate increases rapidly (Warriss, 2000).

Three stages can be described, i.e., initiation, propagation and termination:

i. *Initiation* stage

The hydrocarbon long chain of FA (RH), exposed to light, breaks a C-H bond. The reaction produces a free carbon radical (R°) and an atomic hydrogen (H°):

$$RH \rightarrow R^\circ + H^\circ$$

All free radicals are very reactive substances and they result particularly unstable. The long chain of FA may also react with oxygen to produce a lipid peroxy radical (ROO°):

$$RH + O_2 \rightarrow ROO^\circ + H^\circ$$

Normally the hydrogen involved in initiation stage is located in a α methylene group (-CH_2-) with respect to a double bond (adjacent) (Wood et al., 2007). Therefore, the more double bonds in the fatty acid molecule, the more prone it is to oxidation.

ii. *Propagation* stage

The obtained free radical can react with biatomic oxygen and the reaction produces a lipid peroxy radical:

$$R^\circ + O_2 \rightarrow ROO^\circ$$

This then reacts with another fatty acid molecule to give a lipid hydroperoxide and another free radical:

$$ROO^\circ + RH \rightarrow ROOH^\circ + R^\circ$$

Hydroperoxides are not stable compounds. They normally decompose to give various low weight alcohols, aldehydes and ketones. In particular, the volatile aldehydes obtained in this way, represent the characteristic odours and flavors associated with lipid oxidation.

iii. *Termination*, stage

When two free radicals react together, they destroy each other and they produce a single bond (C-C or C-H). Indeed, a free radical may react with a lipid peroxy radical. The free radicals may also be destroyed by reaction with antioxidant or other molecules, thus creating a huge amount of different stable compounds.

The initiation stage can be catalyzed by a number of factors. Among these are light and heat, metal ions like iron and copper, and the iron-containing haem pigments like myoglobin and cytochromes.

The color of the fat, the attitude to conservation rather than the development of particular odors depends on the degree of oxidation of fatty acids, and then by their level of saturation. In particular by progressive oxidation stages, the adipose tissue (subcutaneous or intramuscular) are exposed to color changing from white to yellow, with the development of unpleasant odors that, at times, make the product unfit for consumption. This means that adipose tissues rich in PUFA will certainly be more desirable from a nutritional point of view, but they show some risks from the technological point of view (Wood et al., 2003).

The technological quality of the swine adipose tissue is mainly represented by its consistency and its resistance against oxidative processes that render it suitable for processing and storage. The consistency is closely correlated to the lipid and water content, to the texture of the connective tissue and to the nature of the fatty acids which constitute the lipids. A low lipid content associated with a high water content leads to a poor consistency of the adipose tissue (Lebret and Mourot, 1988). The fatty acid composition also exerts a key role in

determining the consistency of the adipose tissue: a higher degree of unsaturation corresponds to a lower melting point of the fat and, consequently, to a lower consistency (Enser, 1983).

Furthermore, an adipose tissue excessively rich in unsaturated fatty acids can create serious problems from the technological point of view, since it can easily undergo hydrolytic and oxidative phenomena during manufacturing (Wood et al., 2008). As pointed out by (Lebret and Mourot, 1988), a high organoleptic and technologic quality is not generally associated to a high content of polyunsaturated fatty acids.

The meat transformation industry, however, has different requirements, sometimes opposite to those of the consumers, because the fat plays a very important function in the processes of industrial processing and it is not possible to completely eliminate it. For these reasons, food industries and researches realized a deal with two opposite requests: the first consisted in the consumers opinion and the latter consisted in the processing industry needs.

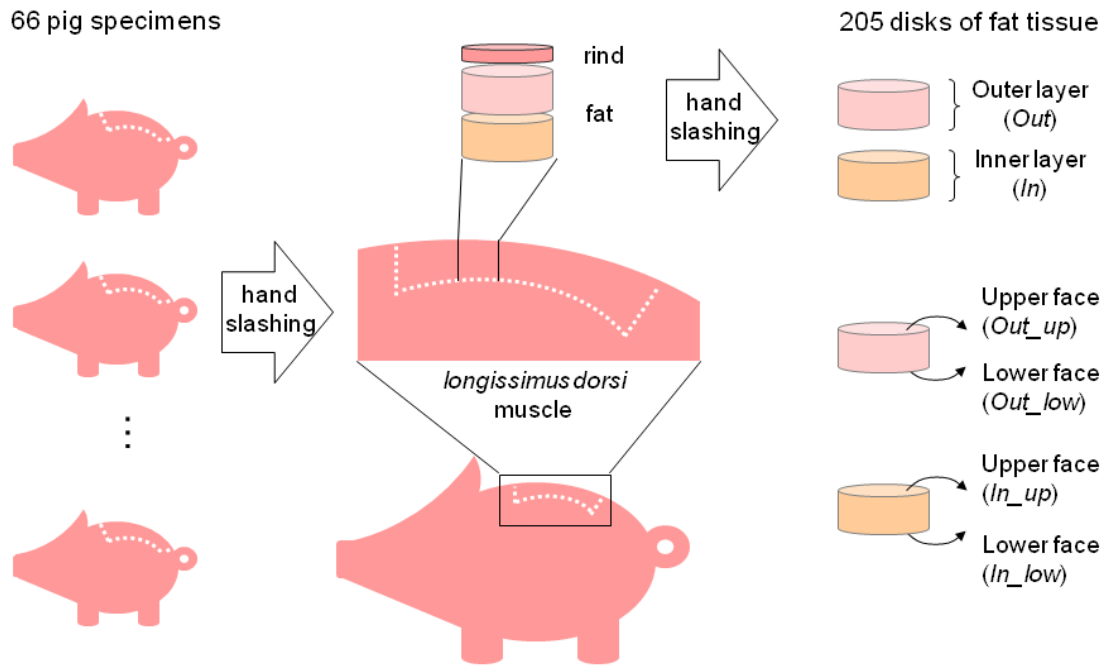
5.3 Classification of swine fat samples from different subcutaneous layers by means of fast and non destructive techniques

5.3.1 Materials and Methods

Samples and analytical methods

Sixty-six pigs from Italian Landrace × Large White crossbreeds provided 205 samples of fat tissue by means of the following sampling procedure. The subcutaneous adipose tissue was hand-slashed by an expert operator at the last rib level in a way to obtain disks of fat tissue having diameter of about 3 cm and thickness ranging from 3 mm to 2 cm. These fat samples consisted in two adjacent layers, lying at different depths with respect to the rind. For classification aim the layer close to the rind (that was previously removed) was labeled as Outer (*Out*) and the layer far from the rind as Inner (*In*). The two layers were then separated by means of a manual cut, after a visual assessment of the line of demarcation of the layers, to gain the corresponding *Out* and *In* samples. To keep track of the sampling point where the following analytical measurements were carried out on the fat disk, an additional labeling was introduced: *Out_up* for the upper face of the *Out* layer, *Out_low* for the lower face of the *Out* layer, *In_up* for the upper face of the *In* layer and *In_low* for the lower face of the *In* layer.

The procedure adopted to delimit and label the fat samples is represented in Scheme 5.3.1.1.



Scheme 5.3.1.1. Sampling scheme of the swine fat tissue samples.

All the collected fat samples were stored in dark conditions at -20 °C. Before analysis, the samples were slowly defrosted at 4 °C for 1 h and then at room temperature for other 30 min. All the measurements were performed at room temperature. Each day, only the samples to be analyzed were defrosted and analyzed following a random order. Then, this order was shuffled, repeated measurements were performed on each sample and, at the end of the daily measurement session, the samples were stored again at -20 °C.

Colorimetric measurements

In extreme synthesis the color of a sample can be defined considering the relationship between the amount of light of a given wavelength reflected by the sample and the amount of light reflected from an object perfectly white. The data of a spectrum of reflection contain all the information related the color of the sample, but it must be mathematically processed for the purpose of a numerical description of the color. For this aim the elaborations most useful are those developed by the Commission Internationale de l'Eclairage (CIE). Such method expresses the color (or the light of each wavelength in the visible spectrum) as the result of a mixture of three primary parameters. The system is known as CIE 1976 L*, a*, b* tristimulus system. In this system, presented in Figure 5.3.1.1., the color of an object is defined in a 3 dimensional space, described by means of three orthogonal variables L*, a* and b* (Ohta and Robertson, 2005).

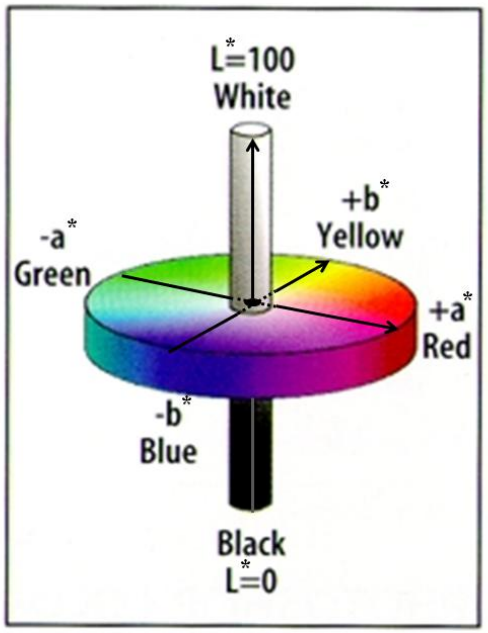


Figure 5.3.1.1. CIE 1976 L*, a*, b* colorimetric space.

The brightness (L*) measures the amount of light. L* goes from black to white with values between 0 and 100, a* and b* are the real chromaticity coordinates. The first parameter (a*) is associated with the color red if positive (0 to +50), and green if negative (from 0 to -50). The second parameter (b*) is related to the yellow for positive values of b* (from 0 to +50), while negative values of b* (from 0 to -50) are related to the blue color. The coordinates a* and b* can be used to express the spectral saturation (*Chroma*) according to equation 5.1:

$$Chroma = \sqrt{(a^{*2} + b^{*2})} \quad (5.1)$$

while the spectral color (*Hue*) is calculated by mean of equation 5.2:

$$Hue = Arctan\left(\frac{b^*}{a^*}\right) \quad (5.2)$$

The routine measurement of color according to the CIE 1976 L*, a*, b* can be performed with cheap instruments known as tristimulus colorimeters. With these instruments, the sample is illuminated by polychromatic light and the reflected light is passed separately through three (sometimes four) filters before reaching a photocell. Based on the characteristics of the source of illumination, the reflected light and filters, the light signals are then processed as tristimulus values.

In this research work the colorimetric measurements were accomplished using a Chroma Meter CR-400 Konica Minolta (CIE standard illuminant D65) tristimulus colorimeter; the standard instrumental procedure for calibration was applied before use.

For each fat disk, four measurements were performed in two different acquisition sessions; in particular, a total of 820 measurements = (205 samples) x (2 acquisitions on the upper and lower faces of the disk) x (2 repeated measurements) have been acquired. A randomized order for the acquisition of repeated measurements was used. Each recorded tern of values, consisting in CIE L*, a* and b* colorimetric parameters, is the result of the instrumental mean of three measurements. In addition, starting from the recorded L*, a* and b* values, two further colorimetric parameters, i.e., the spectral color Hue and the spectral saturation Chroma, have been calculated using the equations 5.1 and 5.2.

The choice of expanding the dataset by including Chroma and Hue is due to the fact that these parameters have often proved to be informative when studying the color characteristics of pork meat and fat (Hallenstvedt et al., 2012; Juárez et al., 2011; Van Oeckel et al., 1999).

FT-NIR spectroscopy

All FT-NIR measurements were performed using a Bruker Optics MPA FT-NIR spectrophotometer equipped with two different sampling tools: Integrating Sphere (IS) working in the region 3800-12500 cm^{-1} and Fibre Optic Probe (FOP) working in the region 4000-12500 cm^{-1}. The spectra were acquired in reflectance mode at 2 cm^{-1} resolution as the average of 64 scans.

For each sampling tool, a total of 1640 spectra has been acquired, as the result of (205 samples) x (2 acquisitions on the upper and lower faces of the disk) x (2 acquisition sessions) x (2 repeated spectra in each session). Also in this case a randomized order for the acquisition of repeated measurements was used.

Data processing and analysis

Explorative analysis: PCA and data organization of colorimetric and FT-NIR data

PCA models were calculated on the autoscaled colorimetric data and on the mean-centered FT-NIR data; the samples lying outside the 99.7% confidence limit in the Q-T^2 plot were identified as outlier and removed from the datasets. Each dataset was split into a training set and into two separate test sets. In particular, the measurements taken on the *Out_up* and the *In_low* faces, i.e., on the extreme faces of each fat sample, whose attribution to the correct layer is sure, were randomly split into the training set (TRN) containing about 2/3 of the objects of each class, and into the first one of the two test sets (TST1), containing the remainder 1/3 of objects. All the other measurements, acquired on the *Out_low* and on the *In_up* faces of the fat layers (i.e., on the adjacent faces of the two different layers) were included in a second test set (TST2). The manual cut of the fat disks involves a degree of uncertainty about the correct cutting position (in addition to the fact that possible changes in the fat composition of each layer within depth may exist), therefore, we preferred to avoid the inclusion of *Out_low* and *In_up* measurements in the model building phase, but to use them only for external validation. As for FT-NIR data, after outlier elimination, the IS and FOP spectra were considered for classification both taking into account the whole spectral range and a limited region of the spectra in the range 10416-5882 cm^{-1}.

The spectral region limited in the interval 10416-5882 cm^{-1} was used to compare the FT-NIR results with the results of another emerging fast and non-destructive technique, i.e., the HyperSpectral Imaging (HSI) analysis (Gowen et al., 2007), that recorded the spectral images of the fat samples in this specific region. The HSI results, however, are not reported in this

thesis because they are part of the work of another Ph.D. student, Carlotta Ferrari. The complete results of this collaborative work are available in (Foca et al., 2013).

PLS-DA classification

All the classification models were computed considering only the two main classes *Out* and *In*. PLS-DA was used as classification method; for each PLS-DA model, the number of Latent Variables (LVs) was chosen on the basis of the minimum value of the Root Mean Square Error in Cross-Validation (RMSECV). In particular, a customized cross-validation (11 deletion groups) was used for the colorimetric and for the FT-NIR data, paying attention to keep the replicate measurements of each sample in the same deletion group. The performance of the classification models was expressed in terms of Efficiency % (see eq. 3.24).

Concerning the used pretreatments, the dataset of colorimetric parameters was autoscaled, while for the FT-NIR spectra, 22 different combinations of pretreatments were tested, since it was not possible to establish in advance the performing better one (Rinnan et al., 2009). In particular, we considered the following pretreatments:

none (N), meancentering (m), first order derivative (d1), second order derivative (d2), linear detrend (det1), quadratic detrend (det2), smoothing (S), Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), that were tested both separately and in the following combinations: d1 + m, d2 + m, det1 + m, det2 + m, S + m, SNV + m, MSC + m, d1 + S + m, d2 + S + m, det1 + S + m, det2 + S + m, SNV + S + m, MSC + S + m.

For each dataset, the model showing the higher EFF value in cross-validation was selected as the best one.

Variable selection

Since the chemical information contained in a NIR spectrum is redundant, it may be often convenient to apply variable selection methods both to decrease the computational load and to increase the robustness of the prediction models. In this work, variable selection models were calculated by means of two different algorithms: iPLS-DA and WPTER.

The implementation of Interval Partial Least Squares regression for Discriminant Analysis (iPLS-DA) was applied to the FT-NIR spectral datasets, in order to select only those intervals containing the most relevant information for classification. Concerning the whole FT-NIR spectra, three different interval widths, i.e., 50, 100 and 200 variables were tested in both *forward* and *reverse* modes. Moreover, iPLS-DA was also applied to both the FOP and IS datasets considering only the restricted 10416-5882 cm^{-1} range; in this case, only one interval width (50 variables) was considered. The pretreatment method as well as the cross-

validation method were the same as those previously considered for the PLS-DA models calculation on the whole spectral range.

The aim of the Wavelet Packet Transform for Efficient pattern Recognition (WPTER) algorithm (Cocchi et al., 2001; Antonelli et al., 2004; Cocchi et al., 2004; Cocchi et al., 2005; Foca et al., 2013) essentially consists in finding a limited number of variables, called wavelet coefficients, which lead to an efficient separation among objects belonging to different classes, through the decomposition into the Wavelet Packet Transform (WPT) domain of the monodimensional signals (e.g., NIR spectra) that describe each object (Cocchi et al., 2001; Foca et al., 2009; Ulrici et al., 2008). A number of parameters can be set on WPTER algorithm to obtain a model, such as the wavelet filter, the decomposition level in the WPT domain and the percentage of wavelet coefficients to be retained.

In the present work, 9 different wavelet filters (db1, db2, sym4, sym5, sym6, sym7, sym8, coif1, coif5) and 5 percentages of preselected wavelet coefficients (0.1%, 0.5%, 1%, 5%, 10%) were used, setting the maximum decomposition level equal to 5. The combination of all these parameters gave a total of 45 cycles of calculation, that were tested on all the four considered FT-NIR datasets (FOP whole spectrum and restricted range, IS whole spectrum and restricted range).

The wavelet coefficients selected by WPTER in each model can be used as input variables for the calculation of discriminant models using internal validation such as cross-validation. For comparison purposes, the same method adopted to classify the whole FT-NIR spectra was used, i.e., PLS-DA calculated on autoscaled wavelet coefficients with customized cross-validation vector (11 deletion groups).

5.3.2 Results and discussion

Colorimetric data

Explorative data analysis

The PCA analysis, performed on autoscaled data (4 PCs and 99.9% explained variance), highlighted the presence of 22 outlier measurements (less than the 3% of the whole dataset), that have been removed before classification. Multivariate PCA analysis has been used as explorative tool. The PC1-PC3 biplot obtained on colorimetric parameters is reported in Figure 5.3.2.1. (model on autoscaled data, 3 PCs selected, 99.4% explained variance). In the considered space of the Principal Components the samples belonging to *Out* and *In* classes are not clearly separated, while PC3 seems to be informative about the fat layer analyzed. In

particular, the fat samples from the inner layer generally present higher values of L* compared to those from the outer layer.

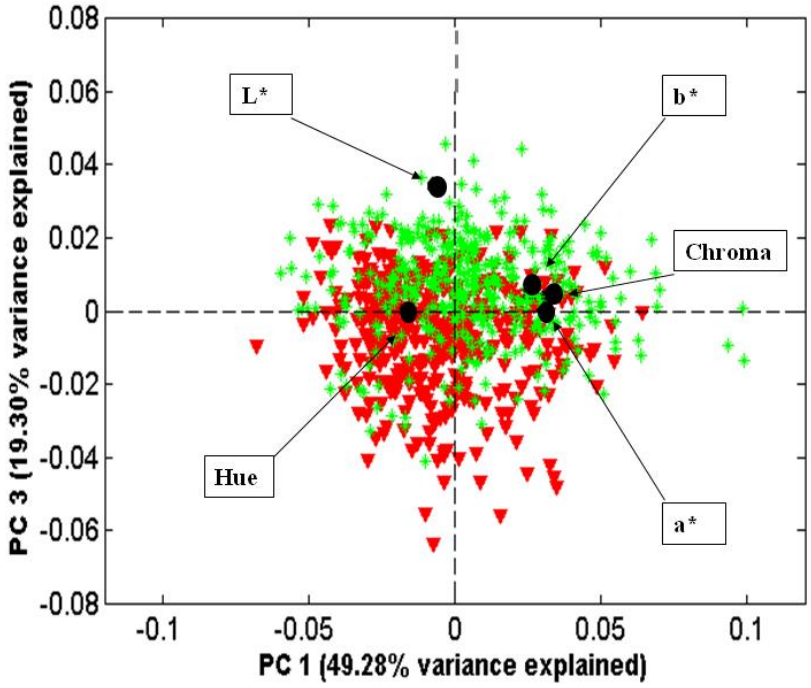


Figure 5.3.2.1. PC1-PC3 biplot obtained on colorimetric data after autoscaling. Class *Out*: red triangles; class *In*: green asterisks.

PLS-DA Classification

After outliers elimination, a training set and two different test sets have been built as previously explained; the number of samples included in each set is reported in Table 5.3.2.1.

	TRN		**TST1**		**TST2**	
	class *Out*	**class *In***	**class *Out***	**class *In***	**class *Out***	**class *In***
Colorimetric data	126	139	63	62	205	203
FT-NIR data (IS)	280	274	125	128	403	408
FT-NIR data (FOP)	284	276	128	128	393	408

Table 5.3.2.1. Number of samples included in the different sets after outliers elimination.

The PLS-DA results obtained on colorimetric parameters are shown in Table 5.3.2.2.

	Colorimetric data
# of variables	5
Pretreatment	Autoscale
# of LVs	3
EFF (TRN)	80.0
EFF (CV)	75.6
EFF (TST1)	**78.1**
EFF (TST2)	**51.9**

Table 5.3.2.2. Results of the best PLS-DA models obtained on colorimetric data for class *Out*.

The PLS-DA model, obtained by using 3 Latent Variables, showed a discrete discrimination capability between samples belonging to the two different fat layers. The Efficiency values obtained for the prediction of the different test sets objects are rather different: for TST1, EFF is equal to 78.1%, while for TST2 is equal to 51.9% only. This results give a first confirmation to our presupposition that *Out_low* and *In_up* samples may have not a certain chemical composition, so that is correct to not include them in the training set for the building of the classification models.

FT-NIR data

Explorative data analysis

The explorative PCA analysis was performed on both the meancentered sets of spectra, i.e., IS and FOP datasets. As for the IS dataset, a 4 PCs model explaining the 97.4 % of the cumulative variance was obtained. In this model, the presence of 22 outliers has been evidenced and the corresponding data have been removed before classification. As for the FOP dataset, the obtained PCA model (4 PCs, 98.7 % explained variance) permitted to remove 23 outliers.

The PC1-PC2 scores plot obtained on the IS dataset is reported in Figure 5.3.2.2., to give a visual representation of the data structure. The reported scores plot evidences that the samples belonging to the two layers are well distinguishable along PC2, though the clusters are visibly superimposed. Similar results have been obtained for FOP dataset: the PC1-PC2 scores plot is useful to have a raw separation of *In* and *Out* samples. In this case, the separation is due to the contemporary contribution of PC1 and PC2. It has to be noticed that both the scores plots showed a slightly higher variability in the samples belonging to the outer fat layer.

A final remark concerns the samples distribution in the scores plots of the two datasets. In the FOP scores plot (Figure 5.3.2.3), the samples are more spread across the PC space with respect to the IS scores plot. This observation suggests that the sampling technique used for spectra collection affects the reproducibility of the measurements. While the fibre optic probe is generally considered a more rapid and flexible tool, the integrating sphere seems to furnish higher acquisition performances in terms of spectra reproducibility.

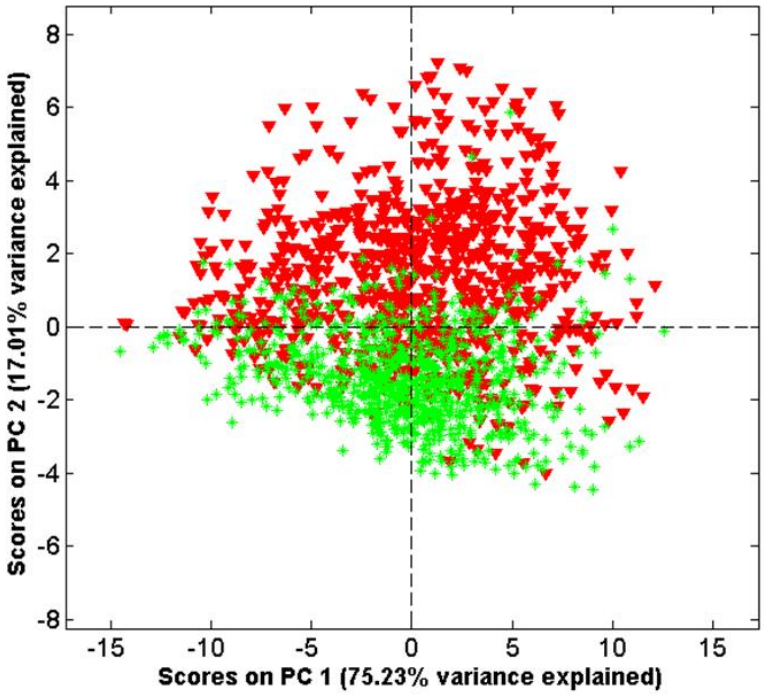


Figure 5.3.2.2. PC1-PC2 scores plot obtained on IS dataset after meancentering. Class *Out*: red triangles; class *In*: green asterisks.

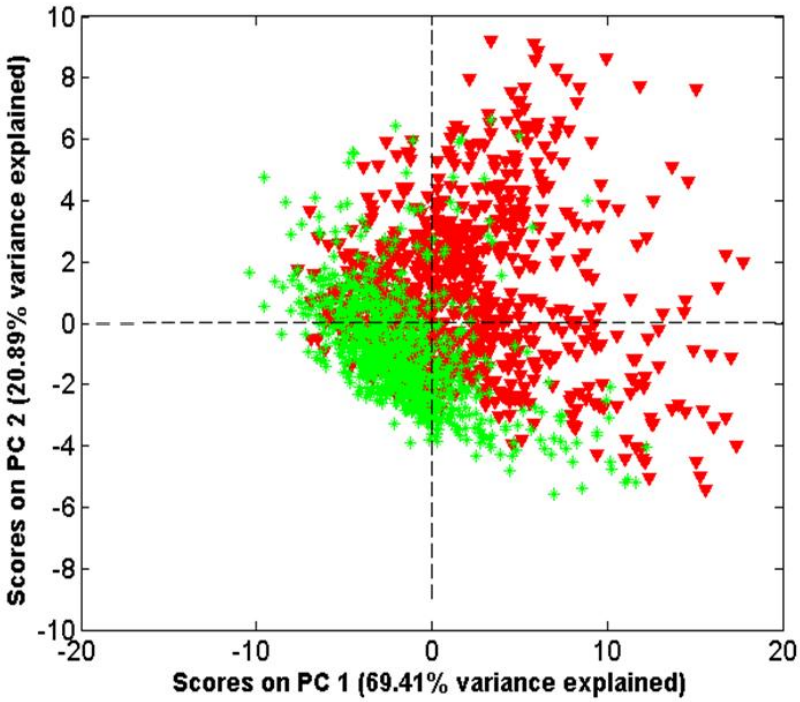


Figure 5.3.2.3. PC1-PC2 scores plot obtained on FOP dataset after meancentering. Class *Out*: red triangles; class *In*: green asterisks.

PLS-DA Classification

Also in this case, after outliers elimination, the training set and the two test sets have been built and the number of samples included in each set is reported in Table 5.3.2.3.

IS dataset (12500-3800 cm^{-1})					
pretreatment	**EFF (CV)**	**#° LVs**	**pretreatment**	**EFF (CV)**	**#° LVs**
N	97.5	12	det1 + m	98.2	12
m	97.3	10	det2 + m	97.7	11
d1	98.6	11	S + m	97.5	10
d2	95.5	11	SNV + m	98.6	12
det1	97.7	12	MSC + m	98.0	11
det2	97.5	12	d1 + S + m	98.6	9
S	97.7	12	d2 + S + m	98.0	9
SNV	98.0	12	det1 + S + m	98.0	10
MSC	98.0	12	det2 + S + m	98.0	12
d1 + m	98.0	10	SNV + S + m	98.7	11
d2 + m	95.8	11	MSC +S + m	98.2	9
FOP dataset (12500-4000 cm^{-1})					
pretreatment	**EFF (CV)**	**#° LVs**	**pretreatment**	**EFF (CV)**	**#° LVs**
N	96.6	8	det1 + m	96.4	7
m	96.8	10	det2 + m	96.6	7
d1	95.7	10	S + m	97.5	11
d2	91.6	10	SNV + m	96.8	7
det1	96.4	8	MSC + m	96.2	7
det2	96.6	6	d1 + S + m	96.8	10
S	97.1	10	d2 + S + m	95.4	10
SNV	96.4	9	det1 + S + m	97.1	10
MSC	96.2	8	det2 + S + m	97.3	8
d1 + m	95.5	9	SNV + S + m	97.1	12
d2 + m	91.9	12	MSC +S + m	97.1	10

Table 5.3.2.3. Comparison of the different pretreatments applied to FT-NIR datasets: performances of the obtained PLS-DA models in terms of Efficiency in cross-validation and dimensionality. The best model for each dataset is highlighted in grey.

In Table 5.3.2.3 the cross-validation Efficiency values obtained with the different pretreatments, together with the number of selected LVs for each model, are reported. As for the models dimensionality, the FOP dataset in general required a lower number of Latent Variables. The Efficiency values, on the contrary, confirm the primacy of the integrating sphere as sampling tool, being the IS results generally higher of the corresponding FOP results. The best models selected for IS and FOP datasets, chosen on the Efficiency in cross-validation, were respectively obtained on the spectra pretreated using a combination of Standard Normal Variate, smoothing and meancentering and on spectra pretreated with smoothing and meancentering.

	FT-NIR data (IS dataset)	**FT-NIR data (IS dataset)**	**FT-NIR data (FOP dataset)**	**FT-NIR data (FOP dataset)**
Spectral range (cm^{-1})	12500-3800	10416-5882	12500-4000	10416-5882
# of variables	4600	2352	4400	2352
Pretreatment	SNV + S + m	MSC + m	S + m	m
# of LVs	11	11	11	12
EFF (TRN)	99.3	100.0	98.8	99.1
EFF (CV)	98.7	99.5	97.5	98.0
EFF (TST1)	**97.6**	**99.2**	**96.0**	**96.0**
EFF (TST2)	**66.6**	**61.1**	**46.5**	**55.7**

Table 5.3.2.4. Results of the best PLS-DA models obtained on FT-NIR data for class *Out*.

In Table 5.3.2.4 the performances of the PLS-DA classification models obtained on IS and FOP datasets, both considering the whole spectra and the spectra limited in the region 10416-5882 cm^{-1}, are reported together with the pretreatments used and the number of LVs selected to build the models.

The IS spectra always gave better results with respect to FOP spectra. In particular, the importance of using a sampling tool more reproducible, as the integrating sphere, seems to be more pronounced when the test samples have not a clearly definite (or maybe homogeneous) composition; in fact, TST2 samples were predicted with an efficiency in the range 61.1-66.6% for IS datasets, while the prediction efficiency reached a maximum value of 55.7% for the FOP datasets. On the contrary, the efficiency in prediction for the TST1 samples was at least equal to 96.0% for all the models, confirming the effectiveness of using FT-NIR spectroscopy to distinguish the fat layers.

Considering the models obtained on the limited range 10416-5882 cm^{-1}, it is interesting to notice that, in prediction, this restricted spectral region gave better results than the whole range with one only exception (EFF of TST2 equal to 61.1% for IS dataset).

Classification after variable selection

Although the results of the PLS-DA classification models can be considered satisfactory, in particular for the samples belonging to the TST1 set, some additional classification models were also computed after the application of different variable selection procedures. Indeed, the removal of noisy and/or uninformative features is often recommended to improve the performances of the classification models, since the presence of these variables may be not only useless for the development of the model itself, but is some cases could even be detrimental (Andersen and Bro, 2010).

As stated before, a number of iPLS-DA models have been computed considering three different sizes for the intervals (only for the whole spectra datasets) and both the *forward* and the *reverse* mode for the intervals selection. In Table 5.3.2.5 only the best model obtained for each dataset has been reported, followed by the number of the selected intervals and variables.

	FT-NIR data (IS dataset)	**FT-NIR data (IS dataset)**	**FT-NIR data (FOP dataset)**	**FT-NIR data (FOP dataset)**
Original spectral range (cm^{-1})	12500-3800	10416-5882	12500-4000	10416-5882
# of original variables	4600	2352	4400	2352
Pretreatment	SNV + S + m	MSC + m	S + m	m
Interval size (variables)	50	50	100	50
***Forward* (F) / *Reverse* (R)**	R	F	R	R
# of selected intervals/variables	34/1700	16/800	38/3800	34/1700
# of LVs	12	10	12	12
EFF (TRN)	100.0	100.0	98.9	99.2
EFF (CV)	100.0	100.0	98.7	98.7
EFF (TST1)	**100.0**	**98.0**	**96.9**	**95.6**
EFF (TST2)	**64.6**	**61.3**	**45.4**	**54.6**

Table 5.3.2.5. Results of the best iPLS-DA models obtained on FT-NIR data for class *Out*.

Three of the best models out of four were obtained using the *reverse* mode for the intervals selection and generally the intervals including a lower number of variables gave better results. As far as the number of selected variables is concerned, two different comments are noteworthy: firstly, the models obtained on the spectra limited in the region 10416-5882 cm^{-1} required about half the variables of the models obtained on the whole spectra; secondly, the models obtained on the IS datasets required about half the variables necessary for the FOP datasets.

With respect to the corresponding PLS-DA models, the models obtained after the interval-based variable selection method showed a slight increase of the performances in cross-validation, while the results in prediction are almost unchanged. Regarding the model dimensionality, no particular differences are observed in the number of latent variables used to build the classification models before and after the variable selection.

In Table 5.3.2.6 the results of the best classification models obtained after the application of the WPTER algorithm as variable selection tool have been reported, together with the parameters used in the selected WPTER cycle, the number of selected WPTER coefficients and the number of LVs used to build the following PLS-DA classification models.

	FT-NIR data (IS dataset)	**FT-NIR data (IS dataset)**	**FT-NIR data (FOPdataset)**	**FT-NIR data (FOPdataset)**
Original spectral range (cm^{-1})	12500-3800	10416-5882	12500-4000	10416-5882
# of variables	4600	2352	4400	2352
Wavelet	db2	sym5	coif5	coif1
% of preselected wavelet coefficients	10%	10%	10%	10%
# of selected WPTER coefficients	14	7	42	58
# of LVs	5	6	4	4
EFF (TRN)	99.1	98.2	98.0	98.0
EFF (CV)	98.6	97.5	97.9	97.9
EFF (TST1)	**97.2**	**97.2**	**94.0**	**91.5**
EFF (TST2)	**61.7**	**55.4**	**54.3**	**52.8**

Table 5.3.2.6. Results of PLS-DA applied on the WPTER coefficients from the best models obtained on FT-NIR data for class *Out*.

For the different datasets, four different wavelet filters were necessary to obtain the best results, while the same percentage of preselected coefficients, i.e., 10%, was used. The IS models were more parsimonious with respect to the FOP models, in fact the number of the selected WPTER coefficients was at the most 14 for the first ones and at least 42 for the second ones. The PLS-DA model dimensionality ranged from 4 to 6 LVs, showing a notable reduction with respect to the models obtained before, both without variable selection and with iPLS-DA variable selection. Comparing the results in Tables 5.3.2.6 and 5.3.2.4 about the models performances, the WPTER algorithm furnished lower values for the prediction efficiency with one only exception, corresponding to the TST2 results obtained for the whole spectra FOP dataset

In a more general discussion about the efficacy of the different variable selection methods applied to the present issue, it cannot be denied that the variable selection did not lead to the improvement of the model results. Maybe, in the particular case of the classification of different fat layers based on FT-NIR analyses, the chemical information spread all over the spectrum is essential to distinguish the samples belonging to the two classes.

Comparison of the spectral regions useful for classification aims

The variable selection methods are by nature conceived to select the portions of a signal that are responsible for the classification of the samples, but also using PLS-DA as classification method is possible to track the signal regions more useful for classification aims by looking at the VIP scores plots. The VIP scores furnish an estimation of the importance of each variable in the projection used in a PLS model: the variables that reach values higher than a fixed limit (usually equal to 1) are considered significant for the model. Hence, the

portions of the spectrum corresponding to significant variables are identified as the spectral regions useful to classification (Chong and Jun, 2005).

In order to gain a possible interpretation of the chemical meaning of the selected variables, in Figure 5.3.2.4. the mean original IS and FOP spectra are reported in comparison with the signal regions selected in the different variable selection/classification models or identified as important for classification in the PLS-DA model without variable selection. In particular, the signal regions in A, B and C refer to the models obtained on IS data and the ones reported in D, E and F to the models obtained on FOP data. Moreover, A and D correspond to the regions above the threshold of 1 in the VIP scores plots, B and E to the intervals selected by iPLS-DA and C and F to the regions selected by WPTER. A similar representation is reported in Figure 5.3.2.5. for the FT-NIR spectra limited in the region 10416-5882 cm^{-1}.

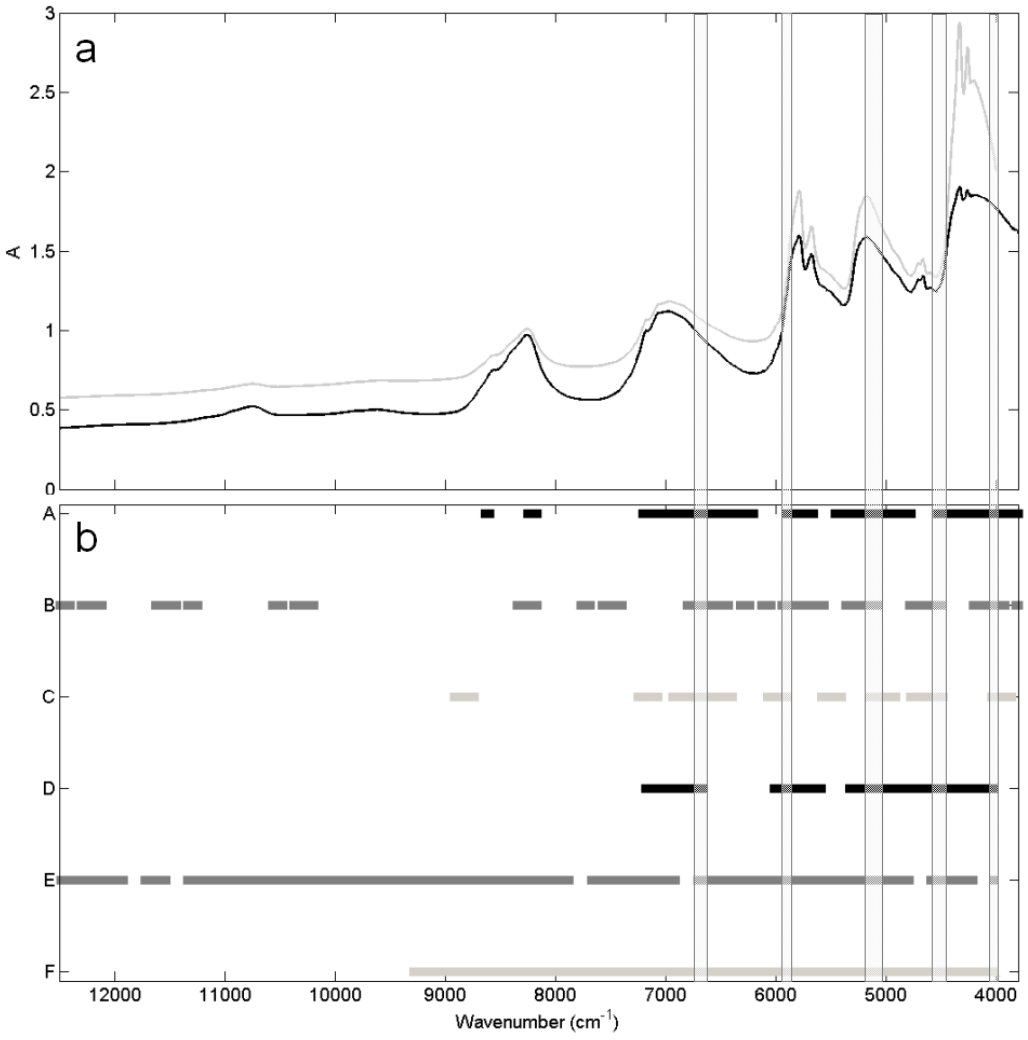


Figure 5.3.2.4. Whole original IS (black) and FOP (grey) mean spectra (a) and signal regions selected in the different models (A-F see explanation along the text). The grey vertical rectangles delimit the regions selected by all the models.

The main absorption bands present in a FT-NIR spectrum acquired on a pig fat sample are principally related to the presence of water and fatty acids. As reported by (Prieto et al., 2009), the O–H bonds are responsible for the broad bands centered at about 6895 and 5155 cm^{-1}, while (Pérez-Juan et al., 2010) attributed to the C–H bonds the peaks around 8250 cm^{-1} (C–H stretch second overtone), 7170 cm^{-1}, 5700 cm^{-1} (C–H stretch first overtone), 4330 cm^{-1} and 4260 cm^{-1} (both combination bands of C–H stretch and deformation). Some regions of the spectrum are indeed informative about the degree of unsaturation in fatty acids carbon chains: the spectral bands in the interval 4545-4345 cm^{-1} are related to unsaturated =C–H and C=C groups (Cozzolino and Murray, 2004) and the band at 5950 cm^{-1} is related to the *cis* unsaturation in carbon chains (Gonzalez-Martin et al., 2003; Gonzalez-Martin et al.,2005). In addition, in the proximity of the water-related bands, in particular in the wide regions 6850-6370 cm^{-1} and 5000-4585 cm^{-1}, also the absorption of the N–H bonds are located (Prieto et al., 2009); the N–H bonds in this kind of samples can be ascribable to the presence of connective tissue.

The most marked evidence in Figure 5.3.2.4. is that the iPLS-DA models considered a higher number of variables as important for classification with respect to the other methods and, in particular, these models included many portions of the signals at wavelengths higher than 10000 cm^{-1}. In general, the different methods made use of most of the spectral region at wavelengths lower than about 7000 cm^{-1}, where many combination bands and the first overtone vibration bands are located. In the figure are highlighted with grey rectangles the five spectral regions where all the best models converged to the same results. In these regions can be found the absorption band related to water at about 5155 cm^{-1}, a part of the spectral region 4545-4345 cm^{-1} related to unsaturated =C–H and C=C groups and a little portion of the 6850-6370 cm^{-1} region where the N–H absorption is located.

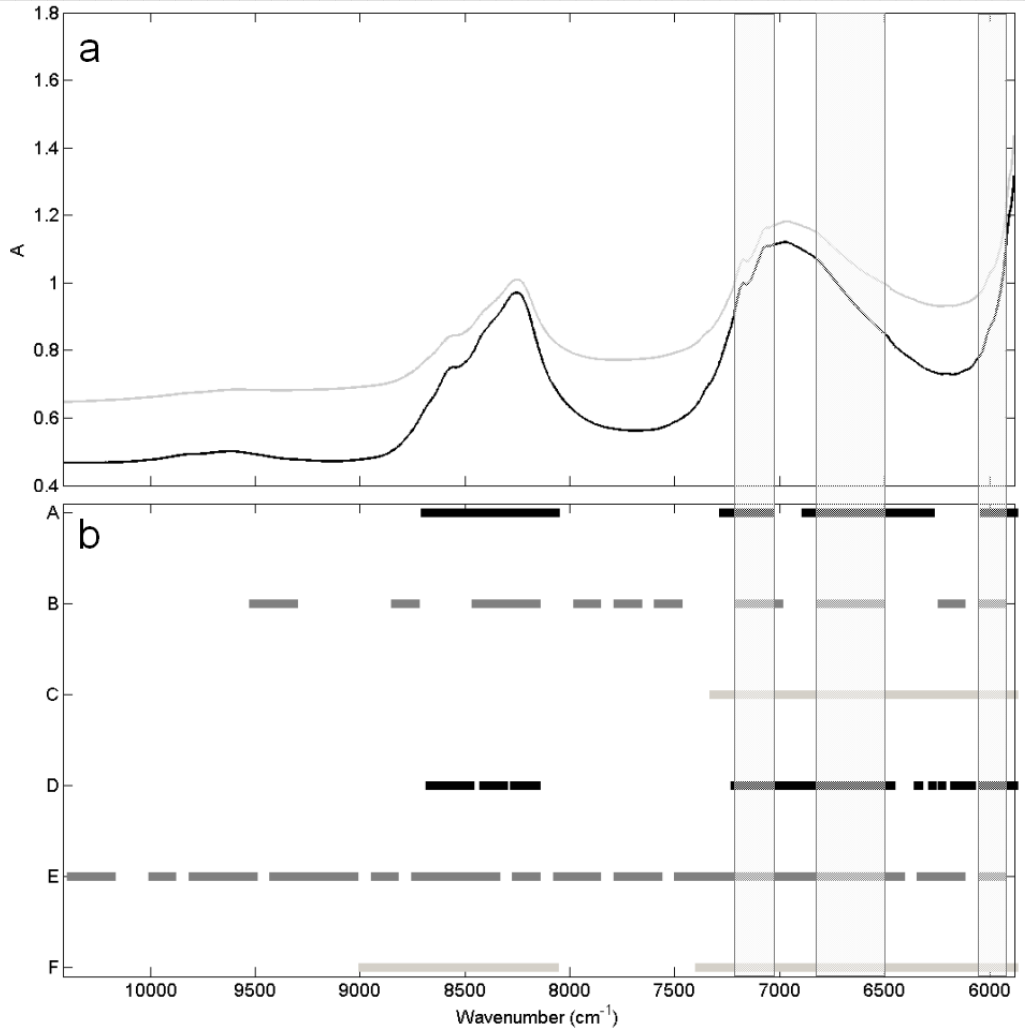
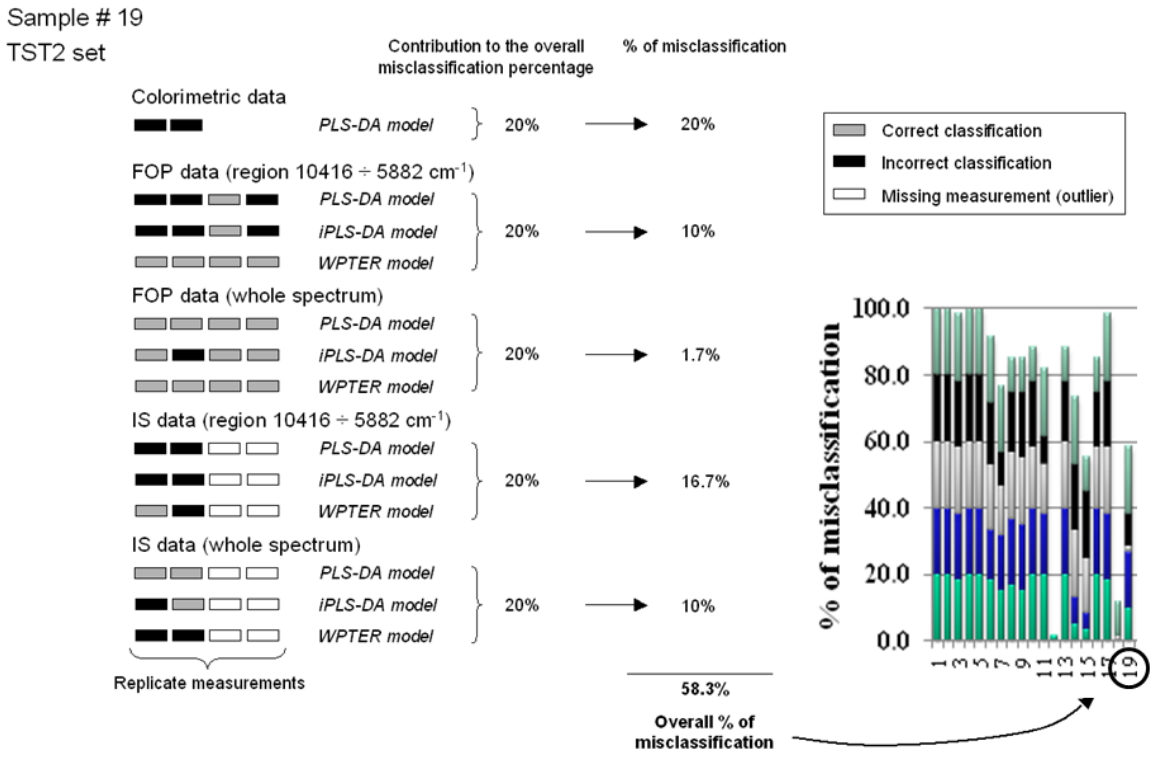


Figure 5.3.2.5. IS (black) and FOP (grey) mean spectra in the range 10416-5882 cm-1 (a) and signal regions selected in the different models (A-F see explanation along the text). The grey vertical rectangles delimit the regions selected by all the models.

As for the models obtained in the region 10416-5882 cm^{-1} (Figure 5.3.2.5.), it can be noticed that also in this case a higher number of variables seem to be important for classification aims in the iPLS-DA models, in particular for the FOP dataset, for which most of the spectral variables have been selected. The VIP scores above the threshold value in the PLS-DA models and the WPTER coefficients selected for FOP data highlighted the importance of the spectral region in the interval 8150-8650 cm^{-1}, where the C–H bonds vibrate in the stretch mode as a result of the second overtone transition. However, the three regions where all the different best models do converge, highlighted with grey rectangles covering the intervals 7195-7050 cm^{-1}, 6805-6520 cm^{-1} and 6025-5945 cm^{-1}, include respectively the regions where the vibrations of the C–H bonds, the N–H bonds and the *cis* unsaturation in carbon chains are located.

Qualitative survey on the misclassified samples

In order to gain an overall survey about the samples that were incorrectly classified in each model, in particular to verify if the same samples have been systematically misclassified by using the different colorimetric and FT-NIR analytical techniques, a proper representation by means of histograms reporting the percentage of misclassification for each fat sample has been obtained (Figure 5.3.2.6.). The building of the histograms followed the procedure described in Scheme 5.3.2.1. for a single sample (i.e., the sample indicated as #19), considering for 'samples' the 410 faces (upper or lower) of the 205 fat disks (inner or outer).



Scheme 5.3.2.1. Procedure to obtain the histograms representing the percentage of misclassification (example on the sample #19 belonging to the TST2 set).

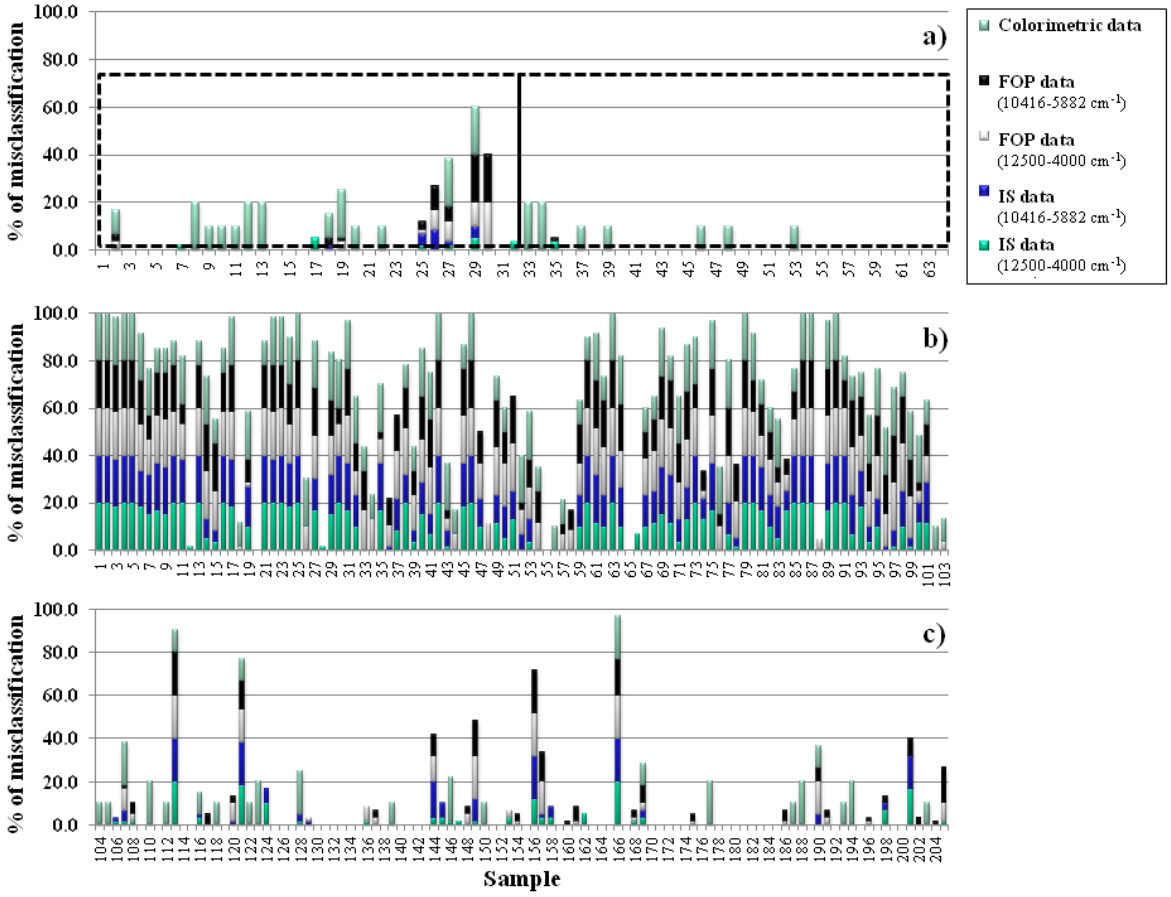


Figure 5.3.2.6. Percentage of misclassification for the fat samples. In (a): TST1 samples, class *Out* samples included in the left dashed rectangle and class *In* samples in the right dashed rectangle. In (b): TST2 samples belonging to the class *Out*. In (c): TST2 samples belonging to the class *In*.

Figure 5.3.2.6. is divided into three parts: part a) reports the percentage of misclassification for each fat sample belonging to the TST1 set (the class *Out* samples are included in the dashed rectangle on the left and the class *In* samples in the dashed rectangle on the right), part b) reports the percentage of misclassification for each fat sample of class *Out* belonging to the TST2 set and part c) the percentage of misclassification for each fat sample of class *In* belonging to the TST2 set

The histogram representation of the TST1 set (Figure 5.3.2.6a), including only the samples at the opposite sides with respect to the cut between the fat layers, i.e., the samples labeled as *Out_up* and *In_low*, put better in evidence with respect to the same results reported in the tables that a relatively low number of samples resulted misclassified and that the colorimetric measurements are the less reliable to the present classification aim. Moreover, a

further and wondering evidence may be noticed: the samples belonging to the class *Out* were misclassified more frequently than the samples of the class *In*.

The same behavior, but much more marked, is observed for the samples of TST2 set which includes the samples neighboring the cut between the fat layers, i.e., *Out_low* and *In_up*. In Figures 5.3.2.6b and 5.3.2.6c a really high degree of misclassification can be noticed for the samples of class *Out* with respect to samples of class *In*, independently from the analytical technique or classification method used to analyze the samples or process the data.

This survey confirms the presence of a real separation surface between the different fat layers, that is reflected in their different physico-chemical characteristics. In particular, the samples corresponding to the outer disks seem to have a greater heterogeneity or, in other terms, a greater variability in fatty acid composition and water amount compared to the inner disks. In addition, the misclassification histograms of the *Out_up* and *Out_low* samples seem to highlight a sort of gradient in the variation of the chemical composition at different distances from the rind. In retrospect, also the explorative PCA analysis gave yet some rough indications, since in the PC1-PC2 scores plots (in Figures 5.3.2.2. and 5.3.2.3.) the class *Out* samples were more widespread in the PC space than the class *In* samples.

5.3.3 Remarks

In this work a comparison among the chemical information bore by two low cost techniques, i.e., tristimulus colorimetry and FT-NIR spectroscopy, is presented with the aim to verify the suitability of these techniques to rapidly discriminate fat samples coming from different subcutaneous layers. In fact, in the Italian industry the different fat layers are generally destined to the manufacturing of different meat products, as a consequence, the availability of cheap, rapid and affordable methods for the characterisation of the overall fat quality is desirable.

The results achieved by colorimetric analysis showed that a relationship between the colorimetric parameters and the characteristics of fat which render it suitable for specific end-uses does exist, even if the classification did not reach excellent results; conversely, the NIR-based spectroscopic methods gave much more satisfactory models. In particular, the results obtained on FT-NIR data showed that the use of the integrating sphere as sampling tool furnished better classification models with respect to the fibre optic probe, probably because of the higher acquisition performances in terms of spectra reproducibility.

In addition, on NIR data some variable selection methods have been used in order to refine the classification results: the model performances did not generally improve, since probably for this particular issue the chemical information spread all over the spectrum is essential to distinguish the samples belonging to the two classes. Anyway, the qualitative interpretation of the chosen spectral regions evidenced the selection of bands related to the vibrations of some chemical bonds typical of a lipid matrix.

A further investigation on the samples that were more frequently misclassified in the colorimetric and FT-NIR models showed a really high degree of misclassification for the samples of class *Out* with respect to samples of class *In*, independently from the analytical technique or classification method used to analyze the samples or process the data. This observation confirms the presence of a real separation surface between the different fat layers, that is reflected in their different physico-chemical characteristics; the samples corresponding to the outer disks seem in fact to have a greater variability in fatty acid composition and water amount compared to the inner disks.

5.4 Application of the FT-NIR spectroscopy for Iodine Value and fatty acids determination on swine fat samples from different subcutaneous layers

5.4.1 Materials and Methods

Samples and analytical methods

Since the FT-NIR spectroscopy is a non-destructive analytical technique, the swine fat samples have been firstly analysed by FT-NIR spectroscopy, then by traditional analyses. For this reason, after the spectra acquisition, aliquots of the same fat samples used for classification purposes (see Scheme 5.3.1.1. and the corresponding explanation), were adopted also for Iodine Value (IV) determination and Gas Chromatographic (GC) analysis.

FT-NIR spectroscopy

The FT-NIR spectra acquired for classification purposes (see *FT-NIR spectroscopy* section in paragraph 5.3.1) have been organized in new datasets, appropriate for the prediction of IV and fatty acids (FAs) composition by means of multivariate calibration techniques. In particular, both for IS and FOP sampling tools, the replicate and repeated spectra acquired on the two faces of each of 205 fat disks (belonging to 66 specimens) have been averaged over the single specimens. In this way, two new datasets composed of 123 mean spectra each, 61 corresponding to class *Out* samples and 63 to class *In* samples, have been obtained. This procedure was necessary to match the number of spectra with the number of classical determinations.

Iodine Value determination

The most common determination of adipose tissue firmness is Iodine Value, that is a measure of the unsaturation of fatty acids included in glycerides compounds (try and diglycerides) and it is expressed in terms of the amount of iodine absorbed by the adipose tissue sample. Basically, IV determines the unsaturation level of the adipose tissue through the number of double bonds in the fatty acids that react with Iodine (I_2). As the swine adipose tissue becomes rich in saturated fatty acids (harder), the iodine value becomes smaller and the

fat is solid and firm; as the swine adipose tissue becomes rich in unsaturated fatty acids (softer), the iodine value becomes greater and the fat is soft and greasy. IV is determined using a direct chemical method, i.e., the Wijs method (AOAC, 1984), starting from a mixture of lipids previously extracted by the Soxhlet extraction method using n-hexane as solvent (IUPAC, 1979). The Soxhlet extraction represents a traditional laboratory analysis able to furnish a separation between water, proteins and lipids contained in the fresh adipose tissue. Normally the extracted mixture of lipids exhibits a high degree of purity and can be used as starting material for other chemical analysis on fat matrix (e.g., IV determination and gas chromatographic analysis).

More in detail, the IV determination, according to Wijs method, represents a case of redox titration (iodometry) in which the iodine I_2 in excess (which have not already react with the double bond of fatty acids) is usually titrated with a standard thiosulfate solution ($Na_2S_2O_3$). In this work, an amount of 0.3 g of extracted lipid mixture is weighed and then it is dissolved in 15 ml of chloroform ($CHCl_3$) by stirring. Afterwards, 25 ml of Wijs reagent, containing I_2, are added to the sample and the solution already obtained is kept in the dark for 1 h. After that, 20 ml of potassium iodide (KI 10% in water) and 100 ml of water are added to the sample solution. The analytical sample solution is titrated with a defined volume of $Na_2S_2O_3$ 0.1 N (ml $Na_2S_2O_3$ sample) using some drops of starch solution as indicator, since it gives an intensely blue complex with iodine; the end is marked by the disappearance of the color indicator.

The redox reaction between iodine and thiosulfate, in which the latter is oxidized to tetrathionate is the following:

$$I_2 + 2\,[S_2O_3]^{2-} \rightarrow 2\,I^- + [S_4O_6]^{2-}$$

This irreversible reaction is fast and quantitative at pH values lower than 7.0.

The same procedure is used for the blank solution which contain no lipids extract.

The iodine value of the sample is finally calculated using equation 5.3.

$$IV = \left[\frac{\left(mL_{\text{Na2S2O3 blank}} - mL_{\text{Na2S2O3 sample}} \right) * 0.1}{sample\ weigh} \right] * 12.69 \quad (5.3)$$

where 0.1 is the concentration of $Na_2S_2O_3$ expressed as equivalent/L and 12.69 is a constant related to the equivalent weigh of iodine.

On the presented fat aliquots, belonging to 66 swine specimens, 158 IV determinations have been obtained. Also in this case, the values of the repeated IV analyses were averaged to gain 123 mean IVs, 61 corresponding to *Out* samples and 63 to *In* samples.

Gas chromatographic analysis

The instrumental technique used to determine the fatty acids composition of the pig fat samples is gas chromatography (Harris, 2005).

To render the fats extracted from swine adipose tissue more volatile, and so suitable for GC analysis, they have to be converted into a mixture of free Fatty Acids Methyl Esters (FAME). As reported by (Minelli et al., 2013), 50 mg of extracted glycerides (IUPAC, 1979) were subjected to trans-esterification and converted in free methyl esters of fatty acids at pH higher than 7.0. The methylation was conducted by means of a methanolic solution of potassium hydroxide (KOH 2N) according to (Ficarra et al., 2010), and adding 100 μL of methyl nonadecanoate (C19:0) as an internal standard.

The so obtained FAME were separated by capillary gas chromatography using a TRACE™ GC Ultra apparatus, equipped with an ultra fast module (UFM) and a fast flame ionization detector.

The GC column adopted (UFM Carbowax column) has the following features: 5 m long, 0.1 mm of internal diameter, and a 0.2 μm thickness for the stationary phase.

The injection of the FAME sample (1 μL) was performed with the split mode technique (splitting degree equal to 1:150).

The carrier gas was helium with a constant flow equal to 0.5 mL/min.

A programmed temperature vaporization injector was set at temperature equal to 240°C and also the detector was constantly kept at 240°C. The column temperature profile used for the analysis is the following:

- starting temperature: 150 °C for 10 sec;
- rate of 102°C/min until 240°C;
- stay at 240 °C for 2.5 min;
- cooling down at room temperature.

The peaks of the FAs were recorded and integrated using the Chrom-Card software (vers. 2.3.3, Thermo Electron Corporation) and identified by comparison with the retention times of standard solutions with known quantities of various methyl esters.

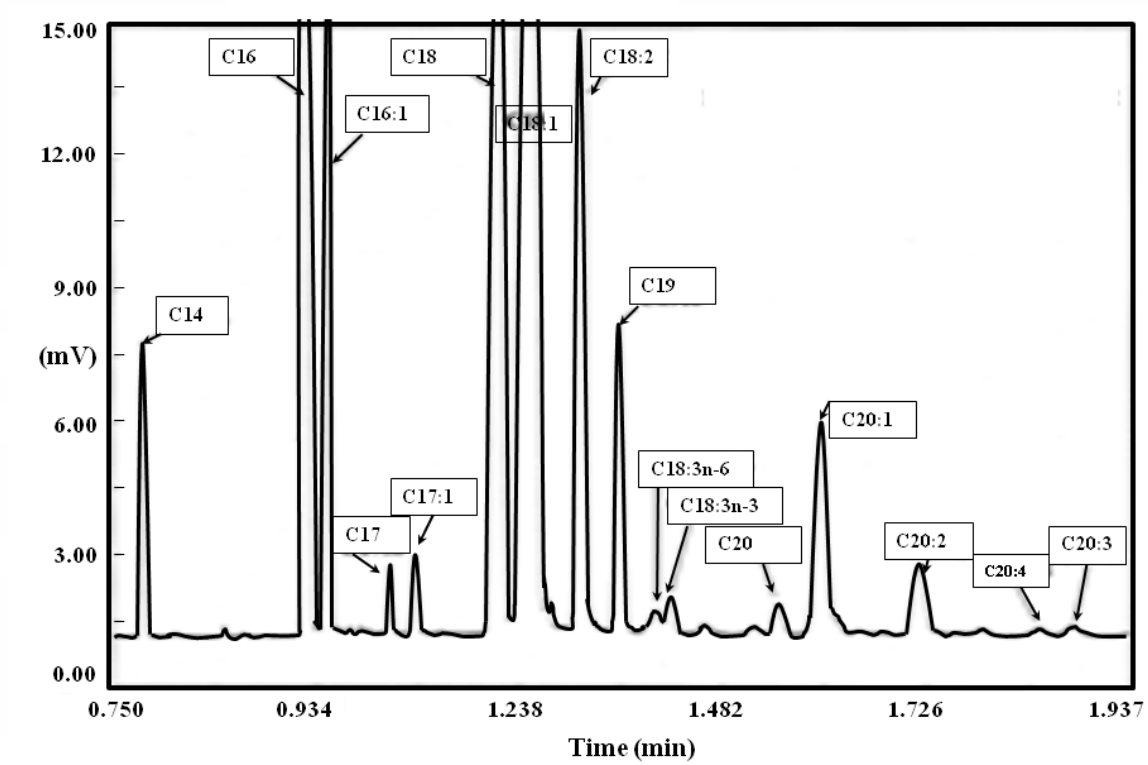


Figure 5.4.1.1. Chromatogram of a swine fat sample.

A visual example of chromatographic profile of a swine subcutaneous fat sample is presented in Figure 5.4.1.1. The peak attribution of the FAs is based on the retention time, according to (Minelli et al., 2013). The peaks corresponding to low carbon chain fatty acids (lower than C14) are not displayed. It can be observed that the longer retention times correspond to the higher weight FAs. Each recorded chromatogram can be used for quantitative purposes using the peak areas ratio between a single peak area and the sum of all peaks areas. Therefore, the amount of each FA in the sample is expressed as FA relative percentage with respect to the total amount of FAs.

According to Figure 5.4.1.1, each GC profile showed the peaks related to seventeen FAs, having carbon chain in the interval C10-C20. As shown in the Table 5.2.2, the highest percentage of FAs in swine meat are mainly related to the presence of C16 (palmitic acid) and C18 (stearic acid) as saturated FAs, and C18:1 (oleic acid) and C18:2 (linoleic acid) as unsaturated FAs. Moreover, (Davenel et al., 1999) and (Geri et al., 1990) confirmed that those FAs are mostly contained in the swine adipose tissue as the tri-acil-glycerols tristearin (glycerol tri C18), triolein (glycerol tri C18:1) and trilinolein (glycerol tri C18:2). Based on this information, it was decided to use only the concentrations of four specific FAs (C16, C18, C18:1 and C18:2) as response variables for the subsequent construction of calibration models.

In addition, starting from the calculated peak areas ratios, three further parameters have been used as response variables, i.e., Saturated Fatty Acids (SFA), Mono Unsaturated Fatty Acids (MUFA), and Poly Unsaturated Fatty Acids (PUFA). These parameters have often proved to be more informative when studying the chemical characteristics of swine meat and fat, compared to a single FA percentage (Piasentier et al., 2009; Lo Fiego et al., 2010; Zamora-Rojas et al., 2013).

SFA, MUFA and PUFA have been calculated as the sums of the percentages of the FAs belonging to the corresponding category:

$$SFA = [\Sigma\ (Saturated\ fatty\ \ Acids\ \%)] = [\ (C_{10}+C_{12}+C_{16}+C_{17}+C_{18}+C_{20})\%] \tag{5.4}$$

$$MUFA = [\Sigma\ (Mono\ Unsaturated\ fatty\ \ Acids\ \%)] = [\ (C_{16:1}+C_{17:1}+C_{18:1}+C_{20:1})\%] \tag{5.5}$$

$$PUFA = [\Sigma\ (Poly\ Unsaturated\ fatty\ \ Acids\ \%)] = [\ (C_{18:2}+C_{18:3n-3}+C_{18:3n-6}+C_{20:2}+C_{20:3}+C_{20:4})\%] \tag{5.6}$$

Also for GC analysis, replicate and repeated measurements permitted to collect 158 chromatograms. The values obtained for C16 (palmitic acid), C18 (stearic acid), C18:1 (oleic acid), C18:2 (linoleic acid), SFA, MUFA and PUFA from replicate and repeated chromatograms have been averaged to gain a GC dataset composed of 123 mean values, 61 corresponding to *Out* samples and 63 to *In* samples.

Data processing and analysis

Statistical survey on IV and GC data

IV and GC values were combined together into a unique dataset, named IV-GC dataset, composed of 123 independent samples (61 corresponding to class *Out* and 63 to class *In*) and 8 variables (IV, C16, C18, C18:1, C18:2, SFA, MUFA and PUFA) to be used in the data analysis.

A statistical survey on the eight variables included in the IV-GC dataset has then been performed. In particular, the mean, the range, the standard deviation and the Experimental

Root Mean Square Error (RMSE Exp) have been calculated on the samples grouped in three ways: considering all the samples together, the *Out* samples only and the *In* samples only. The *RMSE Exp* can be considered a measure of the reproducibility of an analysis and it is calculated using the equation:

$$RMSE_{Exp} = \sqrt{\left[\Sigma\left(\frac{Variance\ of\ duplicated\ values\ of\ selected\ variable}{Numbers\ of\ couples}\right)\right]} \quad (5.7)$$

This statistical survey was also conducted with the purpose to compare the uncertainty associated to the calibration models with that associated to the reference measurements.

PCA and data organization

PCA was used as unsupervised exploratory technique for both IV-GC dataset and FT-NIR datasets with the aim to detect the presence of possible outliers. The outliers were identified as the samples outside the confidence limits of 95 % in the Q-T^2 plot and were removed from the datasets. Therefore all the datasets has been randomly split into a training set (TRN) containing about 2/3 of the objects and into a test set (TST), containing the remaining 1/3 of objects.

PLS calibration

PLS was applied on FT-NIR datasets to predict the chemical composition of swine fat layers in terms of IV, percentages of single FAs, the cumulative SFA, MUFA and PUFA. The IV-GC dataset, used as response matrix, was autoscaled, while the FT-NIR datasets were subjected to the same 22 pretreatments used in the classification section (section *PLS-DA Classification* in paragraph 5.3.1). The pretreatment showing the smallest RMSECV was chosen as the best one.

For each PLS model, the number of LVs to consider was chosen on the basis of the minimum RMSECV in a way not to exceed 8 LVs as maximum number. Moreover, a random subsets cross-validation method (5 deletion groups, 20 iterations) was used.

The performance of the obtained PLS calibration models were expressed in terms of R^2 CV and R^2 Pred.

Variable selection by iPLS

Interval-PLS in the *forward* mode was applied as variable selection method on the FT-NIR datasets, previously pretreated using the best pretreatment as emerged by the PLS models. In the present iPLS models, the whole FT-NIR spectral range, composed of 4400 variables for FOP dataset and 4600 variables for IS dataset, has been divided into a number of intervals consisting of 200 variables each.

5.4.2 Results and discussion

IV and GC data

Statistical survey on IV and GC data

The results of the statistical survey on IV values and GC data are reported in Table 5.4.2.1.

		Mean	Range	Standard deviation	RMSE Exp
all samples	**C16**	25.05	21.63-29.67	1.45	**0.45**
	C18	14.38	9.46-19.55	1.97	**0.31**
	C18:1	43.17	37.23-47.97	2.32	**0.48**
	C18:2	10.01	6.45-17.16	1.86	**0.20**
	SFA	41.87	35.43-48.33	2.97	**0.50**
	MUFA	46.78	40.06-52.06	2.45	**0.50**
	PUFA	11.34	7.41-19.09	2.02	**0.24**
	IV	63.32	53.83-73.20	4.13	**1.79**
***Out* samples**	**C16**	24.35	21.63-27.15	1.21	**0.49**
	C18	13.10	9.46-16.38	1.41	**0.37**
	C18:1	44.02	40.09-47.97	2.04	**0.56**
	C18:2	10.80	8.56-17.16	1.65	**0.19**
	SFA	39.91	35.43-43.96	2.08	**0.51**
	MUFA	47.87	43.33-52.06	2.06	**0.51**
	PUFA	12.22	9.64-19.09	1.80	**0.26**
	IV	65.93	60.20-73.20	3.03	**0.67**
***In* samples**	**C16**	25.76	23.33-29.67	1.33	**0.39**
	C18	15.69	12.16-19.55	1.56	**0.24**
	C18:1	42.30	37.23-46.71	2.28	**0.39**
	C18:2	9.21	6.45-15.56	1.73	**0.21**
	SFA	43.88	38.62-48.33	2.35	**0.50**
	MUFA	45.67	40.06-50.60	2.33	**0.50**
	PUFA	10.45	7.41-16.75	1.84	**0.21**
	IV	60.64	53.83-70.17	3.32	**2.17**

Table 5.4.2.1. Statistical survey on IV and fatty acids composition of the swine fat samples.

The RMSE Exp represents an index of the degree of variability of duplicated measurements. For each IV-GC variable, the RMSE Exp value resulted lower than the corresponding standard deviation, confirming the good reproducibility of these measurements.

Explorative analysis on IV-GC dataset

PCA was performed on the autoscaled IV-GC dataset (3PCs, 97.0% cumulative variance).

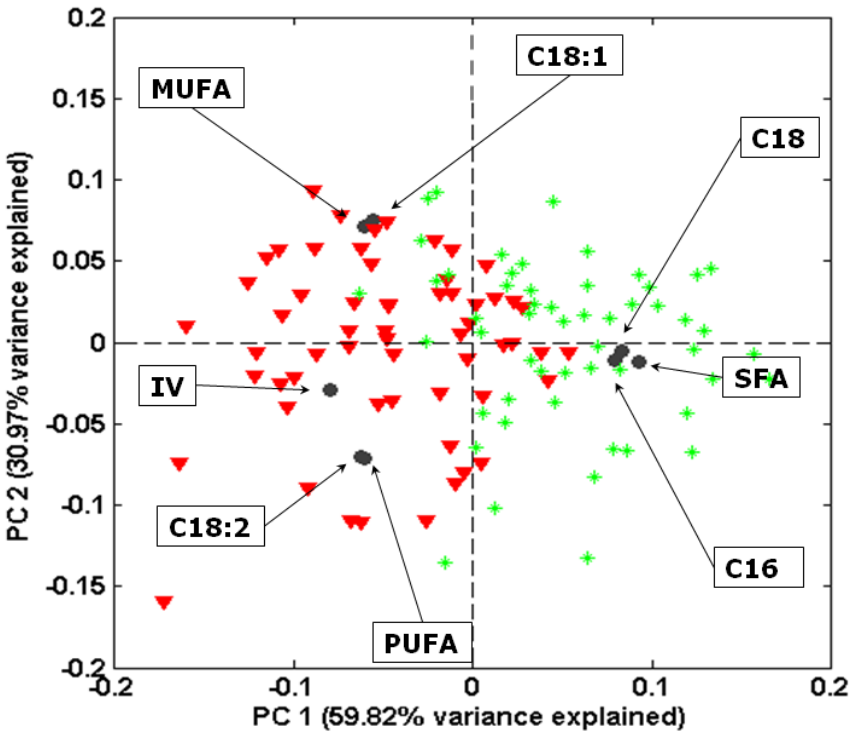


Figure 5.4.2.1. PC1-PC2 biplot obtained on GC dataset after autoscaling. With the exception of IV, the other GC variables are expressed as relative percentage of selected FA. Class *Out*: red triangles, class *In*: green asterisks.

In the PC1-PC2 biplot, shown in Figure 5.4.2.1., the samples belonging to *Out* and *In* classes are somehow separated along PC1, that seems to contain the information about the different chemical composition which characterize the two subcutaneous fat layers.

In particular, the class *Out* fat samples generally present higher values for the variables C18:1, C18:2, PUFA, MUFA and IV, located at high negative values of PC1 in the PCs space. Conversely, the *In* samples present lower values for the variables C16, C18 and SFA, located at positive values of PC1.

FT-NIR data

Explorative analysis on FT-NIR datasets

The explorative PCA analysis was performed on both the meancentered datasets, i.e., IS and FOP datasets. As for the IS and FOP datasets, 3 PCs models explaining the 97.1% and 98.9% of the cumulative variance, respectively, were obtained.

The PC1-PC2 scores plot obtained on the IS dataset is reported in Figure 5.4.2.2., to give a visual representation of the data structure.

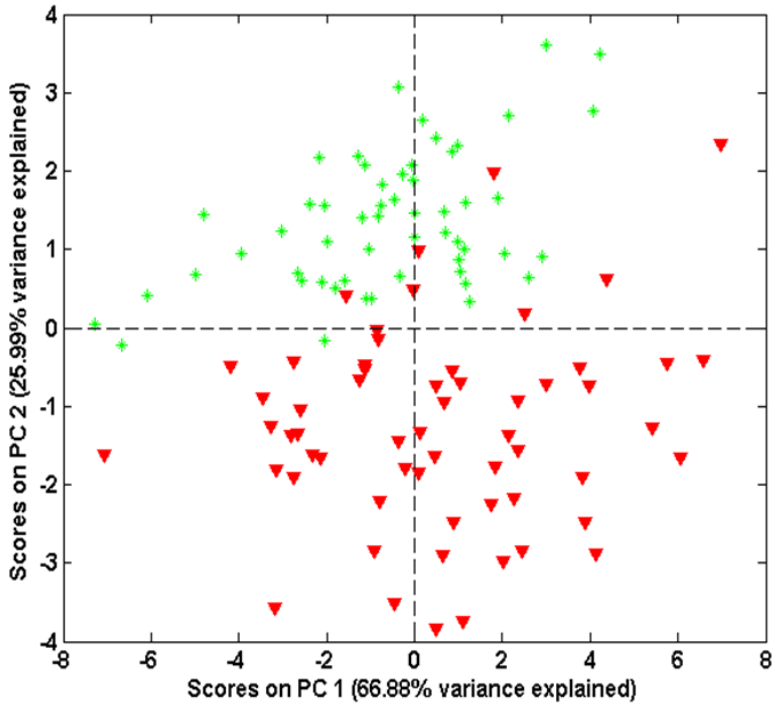


Figure 5.4.2.2. PC1-PC2 score plot obtained on FT-NIR dataset (IS dataset) after meancentering. Class *Out*: red triangles, class *In*: green asterisks.

Also in this case, the reported scores plot shows that the two subcutaneous fat layers are well distinguishable along PC2, though the clusters are visibly superimposed, according to what was reported in paragraph 5.3.2. Similar results have been obtained for FOP dataset (Figure 5.4.2.3.).

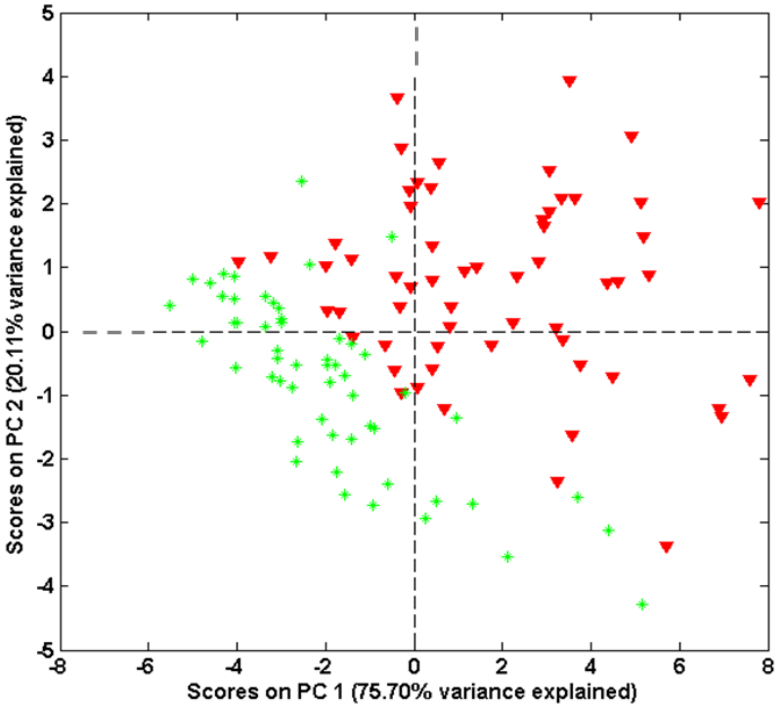


Figure 5.4.2.3. PC1-PC2 score plot obtained on FT-NIR dataset (FOP dataset) after meancentering. Class *Out*: red triangles, class *In*: green asterisks.

In both models, the presence of outlier samples have been evidenced and the corresponding data have been removed before classification.

A qualitative comparison between the samples removed from the datasets used in the calculation of the calibration models was accomplished. It appeared that 5 out of 9 samples defined as outliers and then deleted, showed an anomalous behavior when analyzed with all the used analytical procedures (FT-NIR, IV and GC) . The remaining 4 outlier samples showed an anomalous behavior in only one analytical technique. Hence the deletion of the same 9 objects in all the datasets was necessary to match the datasets before calibration. After the elimination of 9 outliers, the datasets were then composed by 114 independent objects (59 samples corresponding to the outer fat layer and 55 to the inner fat layer); 82 objects were included in the TRN set and 32 in the TST set.

PLS calibration models

In order to predict each response variable, several PLS1 calibration models have been computed.

		IS dataset					
		pretreatment	**LV**	**RMSECV**	**RMSEP**	**R^2 CV**	**R^2 Pred**
all samples	**C16**	d1+S+m	5	1.20	1.02	0.29	0.56
	C18	det1+S+m	8	1.12	1.38	0.63	0.59
	C18:1	MSC+S+m	7	2.31	1.72	0.16	0.49
	C18:2	d2+S+m	8	1.23	0.95	0.57	0.74
	SFA	d2+S+m	7	1.79	1.40	0.64	0.81
	MUFA	det2+S+m	8	2.19	1.58	0.28	0.64
	PUFA	d2+S+m	8	1.29	1.04	0.61	0.75
	IV	d2+S+m	8	2.19	1.68	0.72	0.83
***Out* samples**	**C16**	d1+S+m	4	1.21	1.22	0.15	0.16
	C18	det1+S+m	4	1.17	1.57	0.21	0.06
	C18:1	MSC+S+m	3	2.42	1.97	0.01	0.04
	C18:2	d2+S+m	8	1.42	1.35	0.41	0.44
	SFA	d2+S+m	4	1.68	1.55	0.41	0.70
	MUFA	det2+S+m	4	2.39	1.89	0.01	0.26
	PUFA	d2+S+m	8	1.45	1.41	0.48	0.49
	IV	d2+S+m	8	1.93	2.16	0.64	0.45
***In* samples**	**C16**	d1+S+m	4	1.32	0.91	0.04	0.45
	C18	det1+S+m	3	1.32	1.54	0.08	0.28
	C18:1	MSC+S+m	4	2.40	1.79	0.04	0.27
	C18:2	d2+S+m	8	1.52	1.06	0.19	0.65
	SFA	d2+S+m	8	2.12	1.18	0.21	0.74
	MUFA	det2+S+m	5	2.40	1.74	0.11	0.51
	PUFA	d2+S+m	8	1.58	1.22	0.20	0.59
	IV	d2+S+m	8	3.31	0.97	0.17	0.90

Table 5.4.2.2. Results of the best PLS models obtained on IS dataset.

		FOP dataset					
		pretreatment	**LV**	**RMSECV**	**RMSEP**	**R^2 CV**	**R^2 Pred**
all samples	**C16**	det2	5	1.15	1.12	0.34	0.47
	C18	det1	7	1.13	1.16	0.62	0.71
	C18:1	det1 m	3	2.20	1.97	0.15	0.23
	C18:2	MSC+S+m	8	1.23	0.94	0.57	0.76
	SFA	det2+m	6	1.67	1.33	0.67	0.84
	MUFA	det2+m	8	2.19	1.71	0.32	0.46
	PUFA	det1+m	8	1.34	1.06	0.58	0.75
	IV	det2+S+m	7	2.03	2.04	0.76	0.75
***Out* samples**	**C16**	det2	5	1.21	1.30	0.12	0.11
	C18	det1	6	1.11	1.21	0.31	0.45
	C18:1	det1 m	3	2.28	1.72	0.04	0.16
	C18:2	MSC+S+m	8	1.16	1.00	0.59	0.69
	SFA	det2+m	6	1.64	0.94	0.42	0.88
	MUFA	det2+m	7	2.08	1.60	0.21	0.44
	PUFA	det1+m	8	1.28	0.87	0.60	0.80
	IV	det2+S+m	8	1.68	1.81	0.72	0.62
***In* samples**	**C16**	det2	5	1.26	1.14	0.14	0.22
	C18	det1	2	1.36	1.71	0.00	0.10
	C18:1	det1 m	2	2.31	2.25	0.03	0.04
	C18:2	MSC+S+m	6	1.43	1.21	0.21	0.41
	SFA	det2+m	4	2.04	1.32	0.22	0.66
	MUFA	det2+m	2	2.43	2.16	0.02	0.18
	PUFA	det1+m	6	1.55	1.30	0.21	0.45
	IV	det2+S+m	6	2.58	1.92	0.44	0.62

Table 5.4.2.3. Results of the best PLS models obtained on FOP dataset.

In Tables 5.4.2.2. and 5.4.2.3. an overview on the best PLS models, respectively obtained on IS and FOP datasets, are reported. The Tables also include models dimensionality, mean errors in cross-validation and prediction, and the corresponding correlation coefficients. The models were calculated considering three different datasets, for each sampling tools: the dataset composed by all the samples, the dataset including *Out* samples only and the dataset including *In* samples only.

In general, the calibration performances resulted very poor for C16 and C18:1 variables, while some better results were obtained for the model related to C18 and MUFA, that furnished R^2 Pred values equal to 0.59 and 0.64, respectively, for IS and to 0.71 and 0.46, respectively, for FOP. On the contrary, for the variables C18:2, SFA, PUFA and IV, the TST samples were predicted with a R^2 Pred values at least equal to 0.75.

Although with some exceptions, the PLS models including all the samples gave better performances (higher values of R^2 Pred) compared to the models built considering samples belonging to a specific fat layer. As for the IS dataset, the PLS models built using all the samples generally gave similar results with respect to FOP dataset. On the *Out* samples dataset, the FOP performed better, while on the *In* samples dataset, the better performing sampling tool is the IS.

As for model dimensionality, the PLS models built on samples of class *In* generally required a lower number of LVs with respect to the models built on samples of class *Out*.

Considering the all samples models, only for the IV variable have been obtained PLS models with calculated error in prediction, RMSEP, similar to the experimental error, RMSE Exp. On the contrary, for the other variables the RMSE Exp values range from one-third to one-fifth of the RMSEP values; this evidence confirms that the models are not so effective in predicting the corresponding properties.

These results confirm a certain effectiveness of FT-NIR spectroscopy to predict the iodine value of fat samples without any sample preparation. As for the specific fatty acids, some of them can be satisfactorily predicted, while others are only scarcely predicted.

Calibration performances after variable selection

Although the results of the PLS models can be considered quite satisfactory, in particular for the prediction of the variables C18:2, SFA, PUFA and IV, some additional calibration models were also computed after the application of a variable selection procedure, i.e.,

forward iPLS, this time considering only all the samples together (without any distinction between *In* and *Out* samples).

In Table 5.4.2.4. a summary of the best models after variable selection has been reported, including the model dimensionality in terms of number of LVs, the root mean square errors, the correlation coefficients in cross-validation and in prediction and the number of selected variables and intervals.

	IS dataset						**FOP dataset**					
	# of selected intervals/ variables	**LV**	**RMSE CV**	**RMSE P**	**R^2 CV**	**R^2 Pred**	**# of selected intervals/ variable**	**LV**	**RMSE CV**	**RMSE P**	**R^2 CV**	**R^2 Pred**
C16	1/200	7	0.99	0.96	0.52	0.62	2/400	8	0.83	0.99	0.67	0.60
C18	1/200	8	0.99	0.65	0.71	0.89	1/200	7	0.97	0.76	0.72	0.85
C18:1	2/400	8	1.51	0.99	0.62	0.77	1/200	7	1.55	0.96	0.59	0.80
C18:2	4/800	6	0.94	0.62	0.74	0.89	3/600	7	0.97	0.82	0.72	0.81
SFA	5/1000	8	1.47	1.43	0.75	0.78	2/400	8	1.40	1.61	0.78	0.72
MUFA	1/200	7	1.41	1.11	0.68	0.76	2/400	8	1.52	1.14	0.64	0.75
PUFA	5/1000	6	0.99	0.88	0.75	0.83	2/400	8	1.02	0.86	0.74	0.82
IV	5/1000	5	1.96	1.84	0.77	0.81	2/400	5	1.89	1.72	0.78	0.82

Table 5.4.2.4. Results of the best iPLS models obtained on all the samples together.

The results concerning the iPLS models may be compared with the results of the models obtained on the whole spectra shown in Tables 5.4.2.2. and 5.4.2.3. For most of the modeled variables, the models obtained after the variable selection, showed an increase of the performances in prediction, while the RMSE resulted to be decreased. The improvements in prediction performances affects both the datasets, but are more evident for the FOP dataset. More in detail, the variables C16, C18, C18:1 and MUFA gave R^2 Pred equal to 0.62, 0.89, 0.77 and 0.76, respectively, for IS dataset and 0.60, 0.85, 0.80 and 0.75, respectively, for FOP dataset. Regarding the model dimensionality, after the variable selection a lower number of latent variables have been used to build the calibration models.

Measured-predicted plots and spectral regions useful for calibration aims

Iodine Value

IV was one of the variables satisfactorily predicted by PLS models, for this reason a in-depth analysis on the model details is presented. In Figure 5.4.2.3, IV measured versus IV predicted plots are shown for the IS and FOP complete datasets.

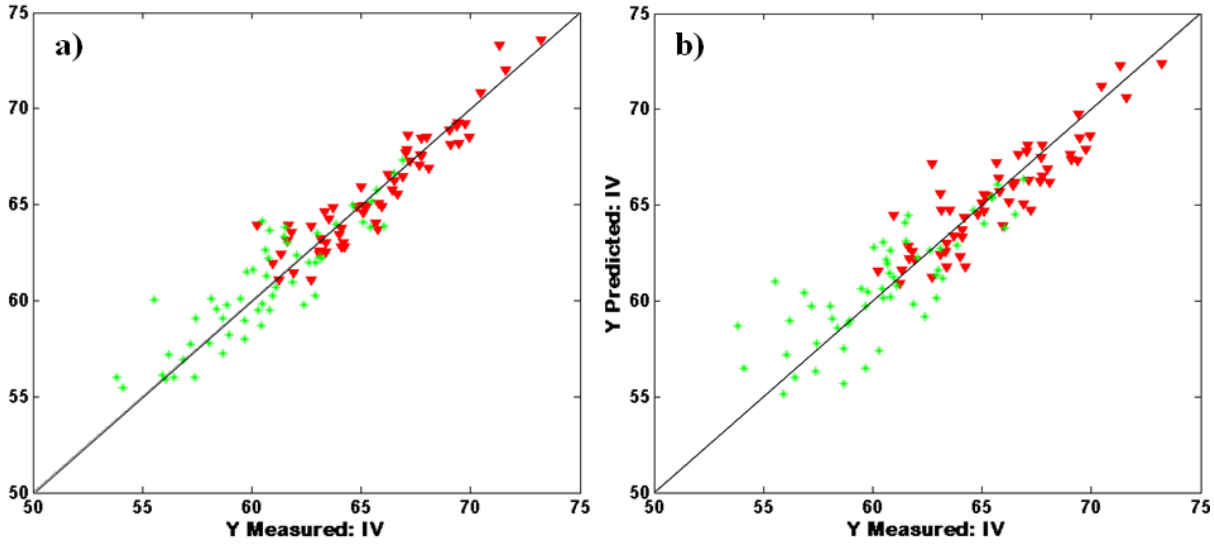


Figure 5.4.2.3. Measured-predicted plots for the PLS models obtained on IS (a) and FOP (b) whole spectra for the variable IV. Class *Out*: red triangles; class *In*: green asterisks.

Figures 5.4.2.3a and 5.4.2.3b show that the sample belonging to different subcutaneous layers present different IV ranges; in particular, the outer fat layer is characterized by fatty acids having a higher number of double bonds.

Concerning the analysis of the spectral regions more useful for calibration aims, a comparison of the spectral variables selected by iPLS and the VIP variables of the corresponding PLS models is reported in Figure 5.4.2.4., together with the original mean spectra of the two datasets (Figure 5.4.2.4a). The signal regions in A and B in Figure 5.4.2.4b, refer to the models obtained on IS data and the ones reported in C and D to the models obtained on FOP data. Moreover, A and C correspond to the regions above the threshold of 1 in the VIP scores plots, B and D to the intervals selected by iPLS.

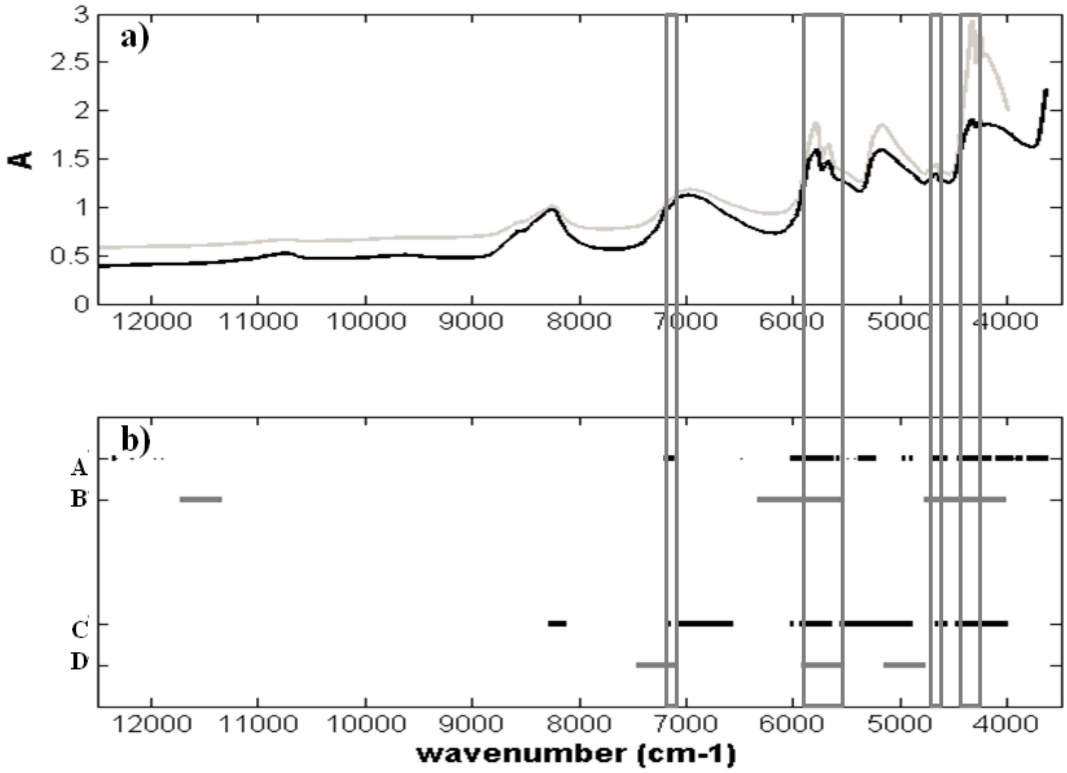


Figure 5.4.2.4. Whole original IS (black) and FOP (grey) mean spectra (a) and signal regions relevant for IV calibration in the different models (A-D see explanation along the text). The grey vertical rectangles delimit the regions selected by all the models.

In agreement with (Hourant et al., 2000), the region between 5500 and 6000 cm^{-1} can be connected with the variation of the IV of fat and oil samples in general. Indeed in that region the first overtone of the C–H stretching, from -CH_2-, -CH_3 and -CH=CH- functional groups of edible swine fats, are located. In the region between 4500 and 4800 cm^{-1} the combination bands of C–H stretching related to *cis* double bonds are located; the intensity of these bands increases with the degree of total unsaturation (Hourant et al., 2000). According to (Li et al., 1999), the absorption bands at 4651 cm^{-1} and 4675 cm^{-1} can be referred to the degree of unsaturation, i.e., the presence of double bonds. The spectral bands in the interval 4545-4345 cm^{-1} are related to unsaturated =C-H and C=C groups (Cozzolino and Murray, 2004) and the band at 5950 cm^{-1} is related to the *cis* double bonds in carbon chains (Gonzalez-Martin et al., 2005).

C18, C18:2, SFA, MUFA and PUFA

As reported by (Ripoche and Guillard, 2001), higher is the degree of homogeneity of the fat samples and higher are the calibration model performances to predict the FAs composition. In this work the fat samples were hand-slashed by an expert operator which has divided the two layers, but this procedure cannot ensure the homogeneity of the fat samples in terms of chemical composition (Foca et al., 2013). Probably due to an intrinsic non-homogeneity of the fat samples, the PLS spectra based models calculated for the evaluation of the percentages of stearic (C18) and linoleic (C18:2) acids are encouraging, but show a certain degree of uncertainty.

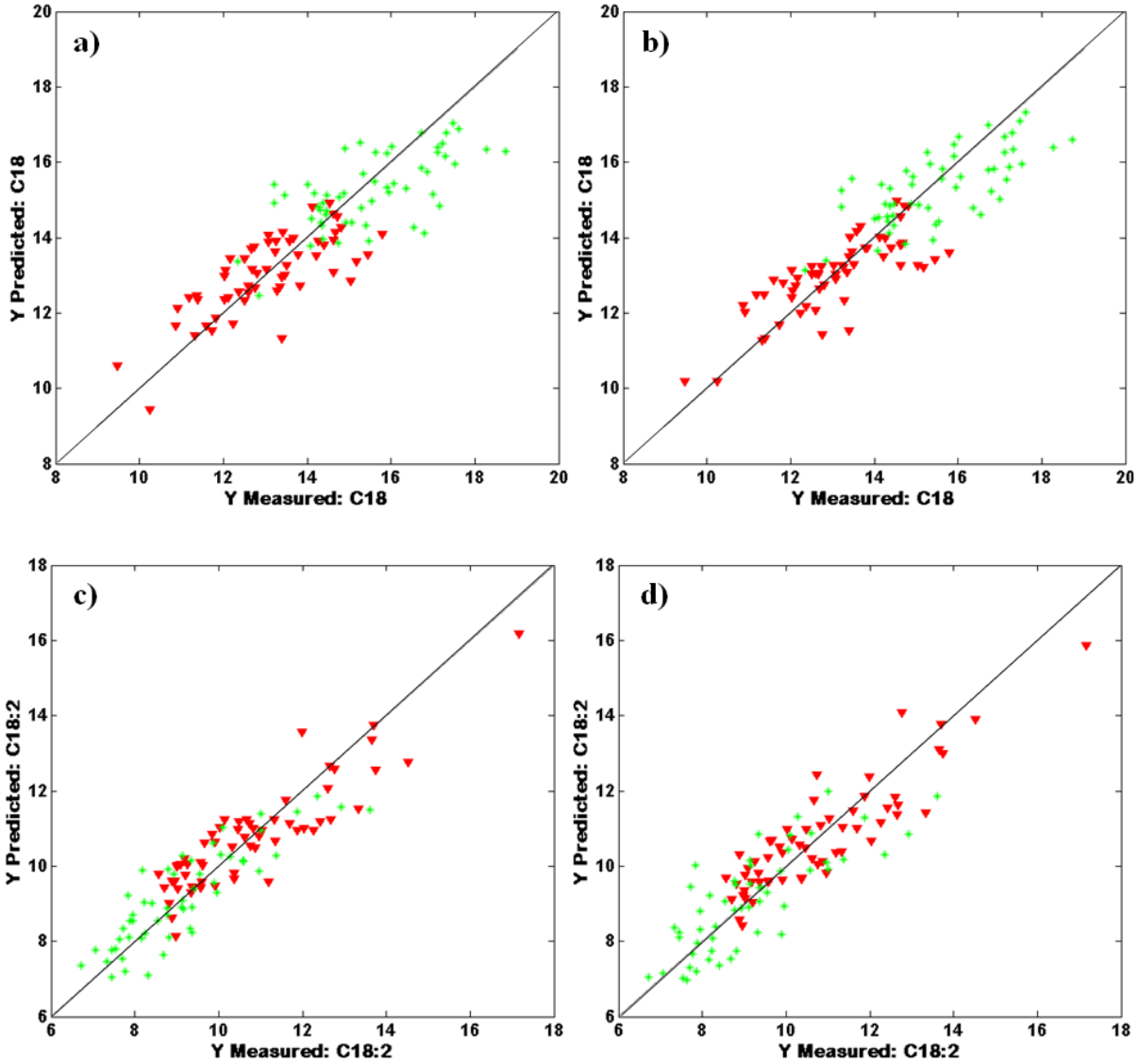


Figure 5.4.2.5. Measured-predicted plots for the PLS models obtained on IS (a and c) and FOP (b and d) whole spectra for the variable C18 (a and b) and C18:2 (c and d). Class *Out*: red triangles; class *In*: green asterisks.

Figure 5.4.2.5. showed the correlation of the variables C18 and C18:2 measured in the laboratory with respect of those predicted by PLS using FT-NIR spectroscopy on the complete IS (Figures 5.4.2.5a and c) and FOP (Figures 5.4.2.5b and d) datasets. Looking at Figures 5.4.2.5a and 5.4.2.5b), i.e., the C18 measured-predicted plots, it can be observed a certain degree of dispersion of the samples around the bisector, according to the obtained R^2 values. The samples belonging to the outer and the inner subcutaneous layers are not so overlapped: the *In* samples seem to generally have greater percentage of stearic acid than *Out* samples. Looking at Figures 5.4.2.5c and 5.4.2.5d, i.e., the C18:2 measured-predicted plots, a slightly lower degree of dispersion of the samples around the bisector is observed. The samples belonging to the outer and the inner subcutaneous layers are mostly overlapped, but the higher C18:2 values are associated to *Out* samples, while the lower C18:2 values are associated to *In* samples.

According to (Azizian and Kramer, 2005), the supply of accurate GC data for the development of effective FT-NIR models is required. In fact, poor correlations between GC and FT-NIR results may be often attributed to misidentified and coeluting analytes in GC analysis. Based on these assumptions, although C18 and C18:2 were fairly well predicted by the PLS models, the cumulative variables SFA, MUFA and PUFA were selected for further investigation of the regions of the NIR spectra related to the overall chemical composition of the subcutaneous fat layers. Indeed, according to equations 5.4, 5.5 and 5.6, those selected variables are indicative of the overall FAs composition of fat samples.

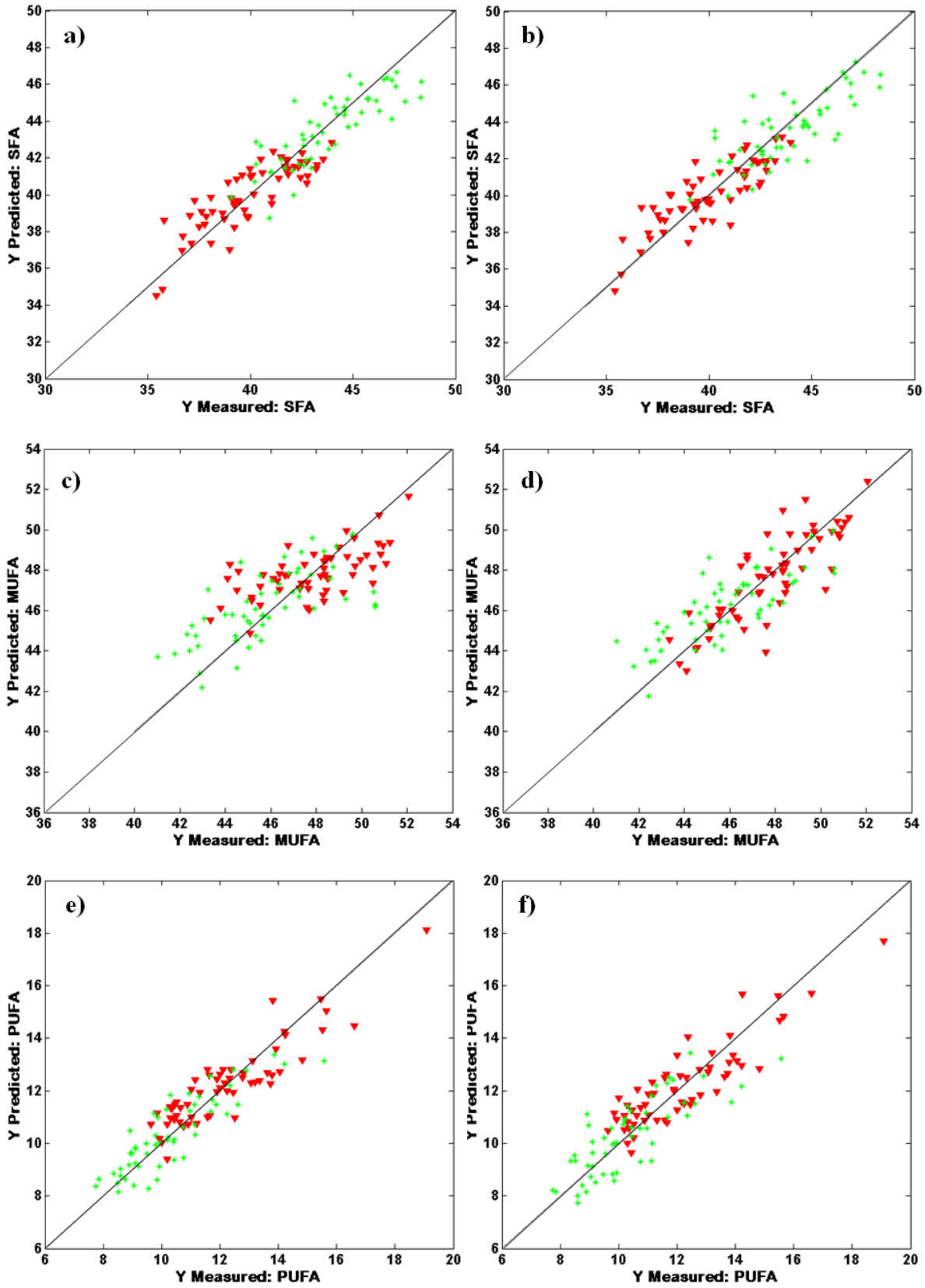


Figure 5.4.2.6. Measured-predicted plots for the PLS models obtained on IS (a, c, e) and FOP (b, d, f) whole spectra for the variables SFA (a, b), MUFA (c, d) and PUFA (e, f). Class *Out*: red triangles; class *In*: green asterisks.

Figure 5.4.2.6. showed the correlation of the variables SFA, MUFA and PUFA measured in the laboratory with respect of those predicted by PLS using FT-NIR spectroscopy on the complete IS (Figures 5.4.2.6a, c and e) and FOP (Figures 5.4.2.6b, d and f) datasets. Looking at Figures 5.4.2.6c and d, i.e., the MUFA measured-predicted plots, it can be observed a certain degree of dispersion of the samples around the bisector, according to the obtained R^2 values in Tables 5.4.2.2 and 5.4.2.3). However, the most interesting remark is that the samples belonging to the outer and the inner subcutaneous layers are mostly overlapped, especially in the cases of variables MUFA and PUFA.

Similarly to what was previously done for the variable IV, even for the variables SFA, MUFA and PUFA a comparison among the spectral regions identified as most relevant in the different calibration models is presented. The wavelength ranges selected by *forward*-iPLS and the VIP variables of the corresponding PLS models are reported in Figure 5.4.2.7, together with the original mean spectra of the two datasets (Figure 5.4.2.7a).

In particular, on the y-axis of the subplots b, c and d, the pair of letters A, B and C, D refer the signal regions selected as important by the models obtained on IS and FOP data, respectively. Moreover, A and C correspond to the regions above the threshold of 1 in the VIP scores plots, B and D to the ranges selected by iPLS.

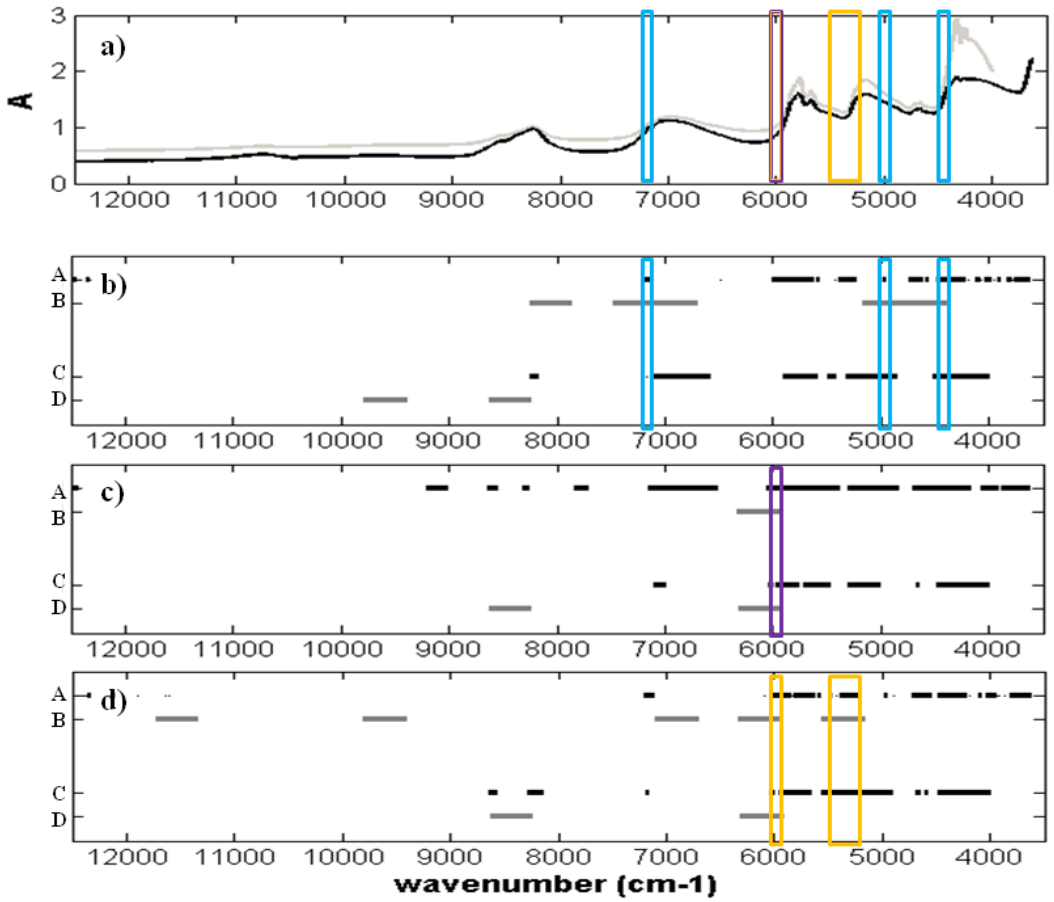


Figure 5.4.2.7. Whole original IS (black) and FOP (grey) mean spectra (a) and signal regions relevant for calibration of SFA(b), MUFA(c) and PUFA(d) (A-D, see explanation along the text). The colored vertical rectangles in subplot (a) delimit the regions selected by all the models.

The FT-NIR calibration models were developed using as reference analyses some determinations performed on heavily treated samples. Band assignments in the NIR region are difficult since a single band may be attributed to several possible combinations of fundamental and overtone vibrations. In the case of swine fat sample, the FT-NIR spectra acquired showed broad peaks related to tri-glycerides and free FAs, but only the free FAs were determined in GC analysis. In particular, the main absorption bands should be related to the presence of tristearin, triolein and trilinolein, three kinds of tri-glycerides which gave rise to the free FAs included in SFA, MUFA and PUFA variables, respectively.

Using a different approach, i.e., fingerprint approach on selected and purified oils (Azizian et al. 2007) and (Azizian and Kramer, 2005), have investigated the main absorption bands of those tri-acil-glycerols. In particular, tristearin is responsible for absorption bands at

5780 and 5800-5850 cm^{-1}, triolein is responsible for absorption bands in the 4500-4750 cm^{-1} and 5600-5900 cm^{-1} regions, while all these tri-acil-glycerols give bands at 5830, 5870 cm^{-1}, 5768 and 5680 cm^{-1} related to the presence of double bonds.

The PLS models obtained in this work are mainly based on the spectral regions at wavenumbers lower than about 7500 cm^{-1} (even lower than 7000 cm^{-1} in most of the cases), where many combination bands and the first overtone vibration bands are located. According to (Pérez-Juan et al., 2010), the subcutaneous swine fat tissues are characterized by the presence of water and fatty acids in different proportion. The broad absorption bands centered at about 6895 and 5155 cm^{-1} can be ascribable to the stretching of the O–H bonds, while the absorption bands attributed to the C–H bonds are located around 8250 cm^{-1} (C–H stretch second overtone), 7170 cm^{-1}, 5700 cm^{-1} (C–H stretch first overtone), 4330 cm^{-1} and 4260 cm^{-1} (both combination bands of C–H stretch and deformation). The SFA calibration models converged in regions mostly concerning the vibrations of the C–H bonds. Conversely, the calibration models related to MUFA and PUFA converged to a narrow interval centered at 6000 cm^{-1}. At 6025-5945 cm^{-1} the combination of the absorption bands related to fibrous proteins (N–H bonds in collagen) and the *cis* unsaturation in carbon chains are located. Finally, also the =C–H and C=C groups present NIR bands in the region 4545-4345 cm^{-1}.

5.4.3 Remarks

The present work is aimed to verify whether the chemical information bore by a fast and non-destructive spectroscopic technique, such as FT-NIR analysis, is able to determine the fatty acid composition and degree of unsaturation of pig fat samples. By now, these parameters are generally determined by using, respectively, gas chromatography and the Wijs analysis, that are definitely more slow and expensive, require the use of a number of chemical reagents and are destructive for the sample.

The calibration models obtained confirm that a relationship between NIR spectra and some chemical characteristics of swine fat does exist. In particular, satisfactory results have been obtained in the prediction of IV, C18:2, SFA and PUFA, for which the correlation coefficients are close to 0.8. Furthermore, a variable selection method has been used on spectroscopic data in order to upgrade the model results: a considerable improvement has been obtained.

The interpretation of the spectral regions selected as informative in the different calibration models obtained, evidenced the typical absorption bands related to the vibrations

of chemical bonds characteristics of saturated and unsaturated triglycerides contained in the subcutaneous swine fat.

As for the characterization of the two subcutaneous fat layers in terms of FAs composition, the obtained results show that the class *Out* samples are characterized by a higher amount of PUFA, a lower amount of SFA and, coherently, higher values of IV; for the class *In* samples the opposite results have been obtained. Indeed, the amount of MUFA seems not to be relevant in fat layers discrimination.

This work also evidenced that the sampling tool used for spectra acquisition plays an important role for spectra reproducibility, confirming that the integrating sphere furnishes more reproducible signals with respect to the fibre optic probe.

5.5 Acknowledgements

The research group of Prof. Lo Fiego (Department of Life Science, University of Modena and Reggio Emilia) is gratefully acknowledged for the samples supply and for the execution of colorimetric, IV and GC analyses.

5.6 References

Andersen, C.M., Bro, R. (2010). Variable selection in regression - A tutorial. *J. Chemom*. 24, 728-737.

Antonelli, A., Cocchi, M., Fava, P., Foca, G., Franchini, G.C., Manzini, D., Ulrici, A. (2004). Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm. *Anal. Chim. Acta*. 515, 3-13.

ASS.I.CA., Associazione Industriali Carne (2007-2012), Annual Reports (2001-2009) and related statistical appendixes.

Ashwell, M., (1993). Diet and heart disease: a round table of factors. London: British Nutrition Foundation.

AOAC., (1984). Official methods of analysis AOAC international, Association of Official Analytical Chemists. Arlington. Official method 28023. Iodine Absorption number Wijs Method.

Azizian, H., Kramer, J.K.G., Winsborough, S., (2007). Factors influencing the fatty acid determination in fats and oils using Fourier transform near-infrared spectroscopy. *Eur. J. Lipid Sci. Technol.* 109, 960-968.

Azizian, H., and Kramer, J.K.G., (2005). A rapid method for the quantification of fatty acids in fats and oils with emphasis on *trans* fatty acids using fourier transform near infrared Spectroscopy (FT-NIR). *Lipids.* 40, 855-867.

Bosch, L., Tor, M., Reixach, J., Estany, J., (2012). Age-related changes in intramuscular and subcutaneous fat content and fatty acid composition in growing pigs using longitudinal data. *Meat Sci.* 91, 358-363.

Bosi, P., Russo, V., (2004). A review: The production of the heavy pig for high quality processed products. *Ital. J. Anim. Sci.* 3, 309-321.

Carrapiso, A.I., Garcia, C., (2005). Instrumental color of Iberian ham subcutaneous fat and lean (biceps femoris): Influence of crossbreeding and rearing system. *Meat Sci.* 71, 284-290.

Chong, I.G., Jun, C.H., (2005). Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* 78, 103-112.

Cocchi, M., Corbellini, M., Foca, G., Lucisano, M., Pagani, M.A., Tassi, L., Ulrici, A. (2005). Classification of breadwheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra. *Anal. Chim. Acta*. 544, 100-107.

Cocchi, M., Foca, G., Lucisano, M., Marchetti, A., Pagani, M. A., Tassi, L., Ulrici, A. (2004). Classification of cereal flours by chemometric analysis of MIR spectra. *J. Agric. Food Chem*. 52, 1062-1067.

Cocchi, M., Seeber, R., Ulrici, A., (2001). WPTER: Wavelet packet transform for efficient pattern recognition of signals. *Chemom. Intell. Lab. Syst*. 57, 97-119.

Cordain, L., Eaton, S.B., Sebastian, A., Mann, N., Lindeberg, S., Watkins, B.A., O'Keefe, J. H., Brand-Miller, J., (2005). Origins and evolution of the Western diet: health implications for the 21st century. *Am. J. Clin. Nutr.* 81, 341-354.

Cox, R., Lebrasseur, J., Michiels, E., Buijs, H., Li, H. Van de Voort, F.R., Ismail, A.A., Sedman, J., (2000). Determination of Iodine Value with a Fourier Transform-Near Infrared Based Global Calibration Using Disposable Vials: An International Collaborative Study. *J. Am. Oil Chem. Soc. (JAOCS)*, Vol. 77, no. 12, 1229-1234.

Cozzolino, D., Murray, I., (2004). Identification of animal meat muscles by visible and near infrared reflectance spectroscopy. *LWT - Food Science and Technology.* 37 (4), 447-452.

Davenel, A., Riaublanc, A., Marchal, P., Gandemer, G., (1999).Quality of pig adipose tissue: relationship between solid fat content and lipid composition. *Meat Sci.* 51, 73-79.

Dell'Orto, V., Sgoifo Rossi, C.A., (2000). Aspetti nutrizionali per la produzione di carne bovina di qualità. *L'Informatore Agrario.* 14, 45-56

Enser, M.B., (1983). The relationship between the composition and the consistency of pig backfat. In Fat Quality in Lean Pigs, *Workshop in the EEC programme*, Brussels, 53-57.

Enser, M.B., (1984). The chemistry, biochemistry and nutritional importance of animal fats. In J. Wiseman (Ed.). Fats in animal nutrition (pp. 23-51). London: Butterworths.

Ficarra, A., Lo Fiego, D.P., Minelli, G., Antonelli, A., (2010). Ultra fast analysis of subcutaneous pork fat. *Food Chem.* 121, 809-814.

Foca, G., Salvo, D., Cino, A., Ferrari, C., Lo Fiego, D. P., Minelli, G., Ulrici, A., (2013). Classification of pig fat samples from different subcutaneous layers by means of fast and non-destructive analytical techniques. *Food Research International.* 52, 185-197.

Foca, G., Cocchi, M., Li Vigni, M., Caramanico, R., Corbellini, M., Ulrici, A., (2009). Different feature selection strategies in the wavelet domain applied to NIR-based quality classification models of bread wheat flours. *Chemom. Intell. Lab. Syst.* 99, 91-100.

Gandemer, G., (2002). Lipids in muscles and adipose tissues, changes during processing and sensory properties of meat products. *Meat Sci.* 62, 309-321.

Geri, G., Zappa, A., Campodoni, G., Franci, O., and Poli, B.M., (1990). Performance: I. Characteristics at 8 days of age and growth rate until slaughter. Relationships between adipose tissue characteristics of newborn pigs and subsequent. *J. Anim Sci.* 68. 1922-1928.

Geri, G., Franci, O., Poli, B.M., Campodoni, G., and Zappa, A., (1990) Performance: II. Carcass traits at 95 and 145 kilograms live weight. Relationships between adipose tissue characteristics of newborn pigs and subsequent. *J. Anim Sci.*, 68, 1929-1935.

Geri, G., Poli, B.M., Zappa, A., Campodoni, G., and Franci, O., (1990). Performance: III. Histological and chemical characteristics of backfat. Relationships between adipose tissue characteristics of newborn pigs and subsequent *J. Anim Sci.* 68:1936-1943.

Girard, J.P., Bout, J., Salort, D., (1988). Lipides et qualités des tissus adipeux et musculaires de porc. Facteurs de variation. I. Lipides et qualités du tissue adipeux. II. Lipides et qualités du tissue musculaire. *Journées de la Recherche Porcine.* 25, 255-278.

Gjerlaug-Enger, E., Kongsro, J., Aass, L., Ødegard, J., Vangen, O., (2011). Prediction of fat quality in pig carcasses by near-infrared spectroscopy. *Animal.* 5 (11), 1829-41

Gonzalez-Martin, I., Gonzalez-Perez, C., Hernandez-Mendez, J., Alvarez-Garcia, N. (2003). Determination of fatty acids in the subcutaneous fat of Iberian breed swine by near infrared spectroscopy (NIRS) with a fibre-optic probe. *Meat Sci.* 65, 713-719.

Gonzalez-Martin, I., Gonzalez-Perez, C., Alvarez-Garcia, N., Gonzalez-Cabrera, J.M., (2005). On-line determination of fatty acids composition in intramuscular fat of Iberian pork loin by NIRs with a remote reflectance fibre optic probe. *Meat Sci. 69,* 243-248.

Gowen, A.A., O' Donnell, C.P., Cullen, P.J., Downey, G., Frias, J.M. (2007). Hyperspectral imaging - An emerging process analytical tool for food quality and safety control. *Trends Food Sci. Tech*. 18, 590-598.

Hallenstvedt, E., Kjos, N.P., Øverland, M., Thomassen, M., (2012). Changes in texture, colour and fatty acid composition of male and female pig shoulder fat due to different dietary fat sources. *Meat Sci.* 90, 519-527.

Harris, D.C., (2005). Chimica Analitica Quantitativa. Zanichelli editore. Seconda edizione italiana condotta sulla Sesta edizione americana.

Hourant, P., Baeten, V., Morales, M.T., Meurens, M., Aparicio, R. (2000). Oil and fat classification by selected bands of near-infrared spectroscopy. *Appl. Spectrosc*. 54, 1168-1174.

INRAN (Revisione 2003). Istituto Nazionale Ricerca Alimenti e Nutrizione, del ministero delle politiche agricole e forestali (www.inran.it). Tabelle di composizione degli alimenti e manuale di sorveglianza nutrizionale e Linee guida per una sana alimentazione italiana.

IUPAC, (1979). Method 1.122. Standard methods for the analysis of oils, fats and derivatives. 6th ed. International Union of Pure and Applied Chemistry, Pergamon Press, New York, NY, USA.

Janz J., Uttaro, B., Robertson, W., (2009). Marbling: consumer acceptance and purchase intent for raw and cooked pork chops of differing marbling levels. Canadian pork Conference. Innovations in Canadian pork quality: An approach towards a differentiation strategy-November 3-4. Montreal, Canada.

Juárez, M., Caine, W.R., Dugan, M.E.R., Hidiroglou, N., Larsen, I.L., Uttaro, B., Aalhus, J.L., (2011). Effects of dry-ageing on pork quality characteristics in different genotypes, *Meat Sci. 88*, 117-121.

Khosla, P., Hayes, K.C., (1994). Cholesterolaemic effects of saturated fatty acids of palm oil. *Food Nutr. Bull.* 15, 119-25.

Lebret, B., Mourot, J., (1988). Caractéristiques et qualité des tissus adipeux chez le porc. Facteurs de variation non génétiques. *INRA Productions Animales.* 11, 131-143.

Li, H., Van de Voort, F.R., Sedman, J., Ismail, A.A., (1999). Rapid determination of *cis* and *trans* content, iodine value, and saponification number of edible oils by Fourier Transform Near Infrared Spectroscopy. *JAOCS.* Vol. 76, no. 4

Lo Fiego, D.P., Macchioni, P., Minelli, G., Santoro, P., (2010). Lipid composition of covering and intramuscular fat in pigs at different slaughter age. *Ital. J. Anim. Sci.* vol.9: e39,

Lo Fiego, D.P., Santoro, P., Macchioni, P., De Leonibus, E., (2005). Influence of genetic type, live weight at slaughter and carcass fatness on fatty acid composition of subcutaneous adipose tissue of raw ham in the heavy pig. *Meat Sci*. 69, 107-114.

Lo Fiego, D.P., Tedeschi, M., Santoro, P., and Nanni Costa, L., (1987). Ricerche sul contenuto di idrossiprolina nel tessuto adiposo del suino pesante. *Atti della Società Italiana delle Scienze Veterinarie.* vol. XLI, Part II, 728-730.

Malmfors, B., Lundstrom, K., Hansson, I., (1978). Fatty acid composition of porcine backfat and muscle lipids as affected by sex, weight and anatomical location. *Swedish J. Agric. Res.* 8, 25-38.

Mayes, P.A., (1996). In: Harper's Biochemistry. 24th Ed. 109-244.

Minelli, G., Macchioni, B., Ielo, M., Santoro, P., Lo Fiego, D.P., (2013). Effects of dietary level of pantothenic acid and sex on carcass, meat quality traits and fatty acid composition of thigh subcutaneous adipose tissue in Italian heavy pigs. *Italian J. Animal Sci*. Volume 12: e52.

Morrissey, P.A., Sheehy, P.J.A., Galvin, K., Kerry, J.P. Buckley, D.J., (1998). Lipid stability in meat and meat products. *Meat Sci.* 49: S73-S86.

Müller, M., Scheeder, M.R.L., (2008). Determination of fatty acid composition and consistency of raw pig fat with near infrared spectroscopy. *J. NIRS.* 16 (3), 305-309.

Murray, C.J.L., Lopoz, A.D., (1997). Mortality by cause for eight regions of the world: global burden of disease study. *Lancet* 349: 1269-76.

Ohta, N., Robertson, A., (2005). Colorimetry: Fundamentals and Applications. John Wiley & sons, Ltd. 130-140.

Pérez-Juan, M., Afseth, N.K., González, J., Díaz, I., Gispert, M., Font Furnols, M., Oliver, M.A., Realini, C.E., (2010). Prediction of fatty acid composition using a NIRS fibre optics probe at two different locations of ham subcutaneous fat. *Food Res. Int.* 43, 1416-1422.

Pérez-Marín, D., De Pedro Sanz, E., Guerrero-Ginel, J.E., Garrido-Varo, A., (2009). A feasibility study on the use of near-infrared spectroscopy for prediction of the fatty acid profile in live Iberian pigs and carcasses. *Meat Sci. 83*, 627-633.

Piasentier, E., Di Bernardo, N., Morgante, M., Sepulcri, A., Vitale, M., (2009). Fatty acids composition of heavy pig back fat in relationship to some animal factors. *Ital. J. Anim. Sci.* vol. 8 (Suppl. 2), 531-533.

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

Ripoche, A., Guillard, A.S., (2001). Determination of fatty acid composition of pork fat by Fourier transform infrared spectroscopy. *Meat Sci.* 58, 299-304.

Rimoldi, G., (2011). Esperienza pratica di gestione crisi nel settore delle carni suine, Associazione Industriali delle Carni (ASS.I.CA), Training Cremona. 6 June.

Rinnan, A., Van den Berg, F., Engelsen, S.B., (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends Anal. Chem.* 28, 1201-1222.

Russo, V., Lo Fiego, D.P., Nanni Costa, L., (1989). Quale suino pesante per l'industria di trasformazione? *Rivista di suinicoltura* n.9.

Scarafoni, A., Magni, C., Duranti, M., (2007). Molecular nutraceutics as a mean to investigate the positive effects of legume seed proteins on human health. *Trends Food Sci. Tech.* 18, 454-463.

Santoro, P. (1984). Fat quality in pig meat with special emphasis on cured and seasoned raw hams. Fat quality in lean pigs. Brussels, 20-21 Sept., 1983, (ed. By Wood, J. D., AFRC Meat Research Institute) Bristol.

SINU. Società Italiana Nutrizione Umana (www.sinu.it), LARN. Livelli di Assunzione di Riferimento di Nutrienti ed energia per la popolazione italiana, riportati nel documento prefinale di sintesi XXXV congresso. Revisione 2012.

Ulrici, A., Cocchi, M., Durante, C., Foca, G., Marchetti, A., Tassi, L., (2008). Multivariate analysis of analytical signals to decipher relevant chemical information. In L. Tassi, & M.P. Colombini (Eds.), New trends in analytical, environmental and cultural heritage chemistry (Chpt. 5). Trivandrum: Research Signpost.

Van Oeckel, M.J., Warnants, N., Boucque, C.V., (1999). Measurement and prediction of pork colour, *Meat Sci.* 52, 347-354.

Warris, P.D., (2000). Meat science - An introductory text. CABI Publishing.

Webb, E.C., O'Neill, H.A., (2008). The animal fat paradox and meat quality: a review. *Meat Sci.* 80, 28-36

WHO/FAO., (Report of 2002). Expert Consultation on Diet, nutrition and prevention of chronic deseases. WHO, Geneva, 28 January, 1 February.

William, E.C., (2000). Importance of n-3 fatty acids in health and deseases. *Am. J. Clin. Nutr*. 71 (suppl):171S-5S.

Wilfart, A., Ferreira, J.M., Mounier, A., Robin, G. Mourot, J., (2004). Effet de différentes teneurs en acides gras n-3 sur les performances de croissance et la qualité nutritionnelle de la viande de porc. *Journées Recherche Porcine*. 36, 195-202.

Wood, J.D., Enser, M., Richardson, R.I., Whittington, F.M., (2007): Fatty acids in meat and meat product. "Fatty acids in foods and their health implications". Third edition. Ching Kuang Chow Ed.

Wood, D., Enser, M., Fisher, A.V., Nute, G.R., Sheard, P.R., Richardson, R.I., Hughes, S.I. Whittington, F.M., (2008). Fat deposition, fatty acid composition and meat quality: a review. *Meat Sci*. 78, 343-358.

Wood, J.D., Richardson, R.I., Nute, G.R., Fisher, A.V., Campo, M.M., Kasapidou, E., Sheard, P.R. Enser, M., (2003). Effects of fatty acids on meat quality: a review. *Meat Sci.* 66, 21-32.

Wood, J.D., Enser, M., (1997). Factors influencing fatty acids in meat and the role of antioxidants in improving meat quality. *Brit. J. Nutr*. 78, Suppl.1, 549-560.

Wood, J.D., (1984). Fat Quality in lean Pigs. Brussels. Commission of the European Communities, 9.

Woodgate, S. Van der Veen, J., (2004). The role of fat processing and rendering in the European Union animal production industry. *Biotechnol. Agron. Soc. Environ.* 8 (4), 283-294.

World Livestock. (2011). - Livestock in food security - Rome, FAO.

Zamora-Rojas, E., Garrido-Varo, A., De Pedro-Sanz, E., Guerrero-Ginel, J.E., Pérez-Marín, D., (2013). Prediction of fatty acids content in pig adipose tissue by near infrared spectroscopy: At-line versus in-situ analysis. *Meat Sci.* 95, 503-511.

Chapter 6: CONCLUSIONS

The results deriving from these three years of work demonstrate that NIR infrared spectroscopy is surely able to bring information about differences in the compositional characteristics of various food matrices. Such differences cannot be generally recognized by a visual inspection of spectra, but coupling this technique with methods of multivariate data analysis, such as classification or regression techniques, it is possible to individuate and, at times, to quantify these differences.

In particular, by using an almost analogous procedure, it has been demonstrated that it is possible to:

i. predict physical, chemical and rheological parameters related to quality of wheat samples analyzed in different physical forms, i.e., as grit, wholemeal flour and white flour;

ii. classify fat samples coming from two different subcutaneous layers to be destined to different end-uses;

iii. quantify in a fast way the iodine value and the fatty acids composition of fat samples.

However, the paramount aim of this thesis work was not to simply resolve specific problems of the food industry, but it was to demonstrate the versatility and the simplicity of use of the proposed method. In addition, other side objectives have been achieved, such as:

- the individuation of the most promising operating conditions - in terms of NIR settings and sampling tools, as well as chemometrics methods and/or approaches - for obtaining the best calibration and classification models;
- the individuation of the spectral regions which are most informative for the different goals of the work, for a possible development of simple and cheap instruments, maybe based on filters and/or lasers, for commercial purposes;
- the achievement of a better understanding of the chemical characteristics and/or the technological behavior of the food samples analyzed;
- the understanding of which analytes are more promising to be actually quantified by means of NIR spectroscopy in real industrial processes.

The procedures here presented could be further developed and in a close future they are likely to be included in the group of methods to be implemented in industry, as already happens for the quantification of moisture and protein content during the execution of the routinely qualitative controls on flours. In some way they could substitute more slow, more difficult to execute and more expensive methodologies.