Universidade do Algarve
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Electrónica e Informática

Development of a data acquisition system for the non-destructive determination of fruit internal quality in an automated calibration line equipped with a visible/near infrared spectrometer

Hélio F. B. Clemente
N.º 29907
Mestrado Integrado em Engenharia Electrónica e Telecomunicações

Faro, 2012

Universidade do Algarve
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Electrónica e Informática

Development of a data acquisition system for the non-destructive determination of fruit internal quality in an automated calibration line equipped with a visible/near infrared spectrometer

Hélio F. B. Clemente
N.º 29907
Mestrado Integrado em Engenharia Electrónica e Telecomunicações

Dissertação orientada por: Doutor Rui Guerra
Professor Auxiliar do Departamento de Física da Universidade do Algarve

Faro, 2012

Abstract

Non-destructive methods for the evaluation of fruit internal quality are of major interest for the agricultural and food processing industry. Until recently the process of evaluating internal quality of fruits was performed manually, involving intensive hand labour and the destruction of fruits. The availability of non-destructive methods for the evaluation of fruit internal quality allowed to speed up the measurements while avoiding fruit destruction, but much of the evaluations are still done manually. Current agro-industrial demands are pushing this type of methods to be more automated, efficient and widespread, in an effort to respond to consumers' demands.

Despite the good results shown in laboratory for non-destructive ºBrix prediction, the application of the method in automatic calibration lines is an enormous challenge. In this project a system was developed to aid in the transfer of the technology from the laboratory to the automated calibration line.

The investigation is focused on non-destructive methods for the prediction of sugar content (ºBrix) on 'Rocha' pears, in automatic calibration lines, using partial least squares regression (PLSR) and diffuse reflectance spectroscopy. All the components of the method were implemented in a software program written in LabVIEW™, including the equipment control and data acquisition, the algorithms of PLSR and the user interface. The system is based on the acquisition of reflectance spectra from the fruits, through a spectrometer, an optical setup consisting of light, fibres and lenses, and a small prototype of an automated grading line. The developed software provides tools for the creation of models to predict the ºBrix and, at the same time, to validate them. Its main goal is to provide rapid design and evaluation of new methods to improve the predictions.

The system was tested by investigating two problems with practical relevance. The first problem was to determine the utility of model recalibration when analysing batches of fruits with characteristics that may differ significantly from those of the fruits used for calibration. The second problem was the quantification of the error induced in the predictions by the random orientation of the fruit in the calibration line. In connection with this question, we have compared the advantages of creating models from fruits with randomized positions versus models created from fruits in aligned positions.

Keywords: reflectance spectroscopy, non-destructive methods for ºBrix prediction, 'Rocha' pear, automatic calibration lines, LabVIEW™

Resumo

A produção mundial de fruta tem vindo a aumentar em todo o mundo nos últimos anos devido à introdução de melhores técnicas agrícolas e de armazenamento, que permitem que culturas sazonais estejam disponíveis todo o ano. Atualmente a procura por frutos continua a aumentar, especialmente na cultura ocidental, onde se procuram hábitos alimentares mais saudáveis, de que os frutos fazem parte integrante. Existe também uma maior preocupação com a qualidade dos frutos por parte do consumidor, a que produtores e cooperativas precisam de responder.

A qualidade é um conceito bastante subjetivo do ponto de vista do consumidor. No entanto, características de cor, tamanho, odor e sabor são com certeza determinantes na sua avaliação. Convém fazer a separação entre propriedades relativas à qualidade externa (cor, tamanho, peso, defeitos) e qualidade interna (teor de açúcares, firmeza, acidez, etc.) Enquanto as qualidades externas dos frutos já são alvo da atenção de produtores e cooperativas, através da integração de balanças e câmaras fotográficas nas linhas de calibração, as características internas ainda são um pouco descuradas por falta de equipamento adequado. Com o aumento da procura de produtos de qualidade, é necessário desenvolver meios para avaliar as características internas dos frutos que possam ser integrados nas linhas de calibração e dar uma resposta rápida e fiável .

Existem formas de avaliar características como o sabor e valor nutricional dos frutos, mas estas técnicas são destrutivas. A avaliação extensiva de um lote de fruta por métodos destrutivos implicaria portanto grandes perdas financeiras, pelo que não é viável. A avaliação destrutiva dos lotes de fruta baseia-se portanto em pequenas amostragens, que podem levar a grandes erros de inferência estatística.

Torna-se assim necessário utilizar técnicas não destrutivas para determinar a qualidade interna do frutos e o processo deverá ser o mais automatizado possível de forma a poder avaliar grandes quantidades de frutos.

Este trabalho é parte de um projeto para determinar o teor de sólidos solúveis (medido em ºBrix) de frutos em linhas automatizadas de calibração através de um método não destrutivo. O projeto surgiu de uma parceria entre o Centro de Electrónica, Optoelectrónica e Telecomunicações (CEOT, Universidade do Algarve), a empresa Calibrafruta (produtores de linhas de calibração automatizadas) e a empresa MCM-Electronics (automação industrial).

O teor de sólidos solúveis afere uma das qualidades internas dos frutos, já que é muito aproximadamente proporcional à concentração de açucares no sumo e portanto é uma medida da doçura dos frutos. A técnica utilizada neste trabalho baseia-se na espectroscopia de refletância difusa na gama do visível e do infravermelho próximo para determinar o ºBrix de forma não destrutiva. Na aplicação específica da espectroscopia de refletância feita neste trabalho, recolhe-se a luz que emerge da polpa do fruto depois de ter sido focada na sua casca. A óptica do sistema é feita de forma a recolher apenas os fotões que entraram dentro da polpa do fruto e não aqueles que foram simplesmente reflectidos pela casca. Assim garante-se que a luz recolhida traz informação sobre o interior do fruto; é esta luz que é analisada pelo espectrómetro, que fornece um espectro de refletância. A informação contida no espectro pode então ser 'convertida' para determinar o ºBrix do fruto.

Para 'converter' esta informação são necessárias métodos estatísticos que relacionem o espectro com o ºBrix. Portanto é necessário conhecer o ºBrix dos frutos, o que, na fase inicial do trabalho, tem de ser feito ainda de forma destrutiva, que habitualmente é a refratometria. É também essencial que se façam muitas medições (de preferência de vários e diversos frutos da mesma variedade) para que se possa inferir uma relação robusta entre os espectros e a qualidade interna dos frutos.

Neste trabalho o método estatístico dos mínimos quadrados parciais (Partial Least Squares - PLS) foi usado para relacionar espectros com os valores de ºBrix de amostras conhecidas, bem como para prever o ºBrix de amostras desconhecidas a partir dos seus espectros. O PLS é um método que determina as direções no hiperespaço das variáveis independentes (neste caso as refletâncias para cada comprimento de onda medido) que simultaneamente melhor explicam a variância dos dados e melhor se correlacionam com as variações correspondentes das variáveis dependentes (neste caso apenas o ºBrix). Usando este método é então possível gerar um modelo que relaciona os espectros com os valores de ºBrix. O modelo pode depois ser aplicado ao espectro de frutos desconhecidos para prever o seu ºBrix.

Como em qualquer regressão, as previsões são sempre afectadas de uma certa incerteza ou erro de previsão. No caso presente esta incerteza é considerável devido à variabilidade biológica das frutos, o que implica correlações relativamente baixas entre os espectros e o ºBrix. O conteúdo bioquímico e estrutural dos frutos varia dentro de parâmetros pouco controlados ou mesmo desconhecidos (por exemplo, é impossível controlar as variações de todos os componentes químicos que podem causar variações nos espectros). Por outro lado, o desempenho de um modelo de previsão pode ir sendo melhorado pela inclusão sucessiva de

mais amostras, de forma a poder simular de forma o mais realista possível o comportamento médio da população. Mas isto também quer dizer que a finitude da amostragem representa em si um factor adicional de erro de previsão.

Já foi comprovado por vários grupos de investigação que é possível usar o método de espectroscopia de refletância para determinar o ºBrix com alguma exatidão (dentro de uma margem de erro aceitável). No entanto aplicar estes métodos numa linha de calibração automatizada (e obter erros aceitáveis) ainda é uma tarefa complexa.

O trabalho que se apresenta neste relatório foi desenvolvido para apoiar a transposição dos métodos descritos acima (usados em laboratório) para uma linha de calibração automatizada. Para isto foi desenvolvido software que permite adquirir espectros, criar modelos e validá-los (uma etapa essencial para verificar a performance de um modelo) em condições que simulam o ambiente real de um calibrador automático industrial.

O software desenvolvido permite controlar o espectrómetro, adquirir e guardar espectros. Os espectros são usados posteriormente (usando o PLS) para criar os modelos que permitem realizar a previsão do ºBrix. O software permite também alterar alguns parâmetros para criação dos modelos. Podem então ser criados e testados diversos tipos de modelos (com parâmetros diferentes) que efetuam previsões ligeiramente diferentes. O objectivo é então gerar um modelo que faça as melhores previsões possíveis (com menor erro). Os testes aos diversos modelos são feitos com base em repetições independentes dos ensaios. O software é construído de forma a gerir de forma simples as repetições de medidas e a estatística associada ao tratamento dos dados decorrentes dessas repetições. Toda a informação é fornecida ao utilizador através de uma interface simples, que lhe permitirá tirar conclusões em tempo real sobre a robustez dos modelos investigados.

Para fazer a prova de teste do software investigaram-se duas questões importantes que surgem na transposição dos modelos do laboratório para a linha de calibração. Os testes foram feitos em pêra 'Rocha'.

A primeira questão surge quando os modelos são criados a partir de um conjunto de frutos com características muito diferentes das características dos frutos para os quais se pretende fazer a previsão (por exemplo, um modelo feito com peras de sequeiro para prever o ºBrix de peras de regadio). Neste caso é comum acontecer que as previsões do modelo apresentem alguma forma de viés relativamente aos valores reais. De uma forma geral, este viés manifesta-se na forma de alguma translação ou rotação da linha de tendência dos resultados esperados. Para eliminar o viés foi testado um método de recalibração dos modelos. A recalibração envolve medir algumas amostras (do lote de frutos que se pretende

realizar a previsão do ºBrix) não destrutivamente (espectros) e destrutivamente (ºBrix). Estes dados são depois usados para fazer o ajuste da previsão ao lote em questão (minimizando o viés).

Os resultados mostraram que a recalibração não é recomendável quando os lotes de calibração e validação são semelhantes. Quando os lotes apresentam diferenças a recalibração pode ser útil. Os testes mostraram que no caso em que o lote de validação é o mais heterogéneo possível, a recalibração pode, por vezes, melhorar os resultados. No entanto, a limitação temporal do trabalho não permitiu recolher peras com características suficientemente distintas para se poder tirar conclusões definitivas.

A segunda questão surge devido à orientação aleatória dos frutos na linha de calibração. Enquanto que no laboratório se posicionam os frutos para se realizar a medição no melhor ponto do fruto (região equatorial), no calibrador a posição é aleatória. Foi demonstrado que a posição aleatória dos frutos faz com que existam flutuações nas previsões. Para minorar este efeito foi testado um método com objectivo de aumentar a robustez dos modelos. Foram medidos os espectros do mesmo fruto em várias posições (realizado em vários frutos), e foram criados modelos usando esses espectros. O modelo criado com espectros repetidos foi comparado com o modelo criada a partir de somente um espectro por fruto. Os resultados demonstram que usar na etapa de calibração múltiplos espectros do mesmo fruto, em várias posições, reduz os erros da previsão do ºBrix.

Palavras-chave: espectroscopia de refletância, métodos não destrutivos de previsão ºBrix, pêra ‘Rocha’, linhas automáticas de calibração, LabVIEW™

Table of Contents

1. Introduction

1.1. Globalization and Fruit Quality

Market globalization has highly increased the demand for more, better and more nutritious fruit. Producers and the fruit processing industry have responded positively by increasing fruit production and availability over the last few decades. This increase is depicted in Figure 1. Globalization requires also a large offer of fruit throughout the year, with reasonable shelf life and quality. In order to achieve this goal, the same fruit varieties/species might be produced in the north and/or south hemispheres, while the seasons change and then properly transported for the consuming markets. Additionally, there is the need for a high storage capacity along the production line, which has been tackled with success through the development of new techniques, which has been shown to grant extended fruit storage life. The increasing consumers' health concerns, particularly in the more developed countries, led to a greater awareness on food safety and nutritive values of fresh commodities. Overall, consumers demand for a large variety of fruit, and these must not only look and taste good, but also exhibit their natural nutritive properties.

Portugal has a great potential to increase fruit production and quality due to its climate and soil conditions, and therefore, increase these commodities exportations.

The production and consumption of fruit are processes out of phase, in terms of both time and space, generating a need to improve market and logistical operations. The typical

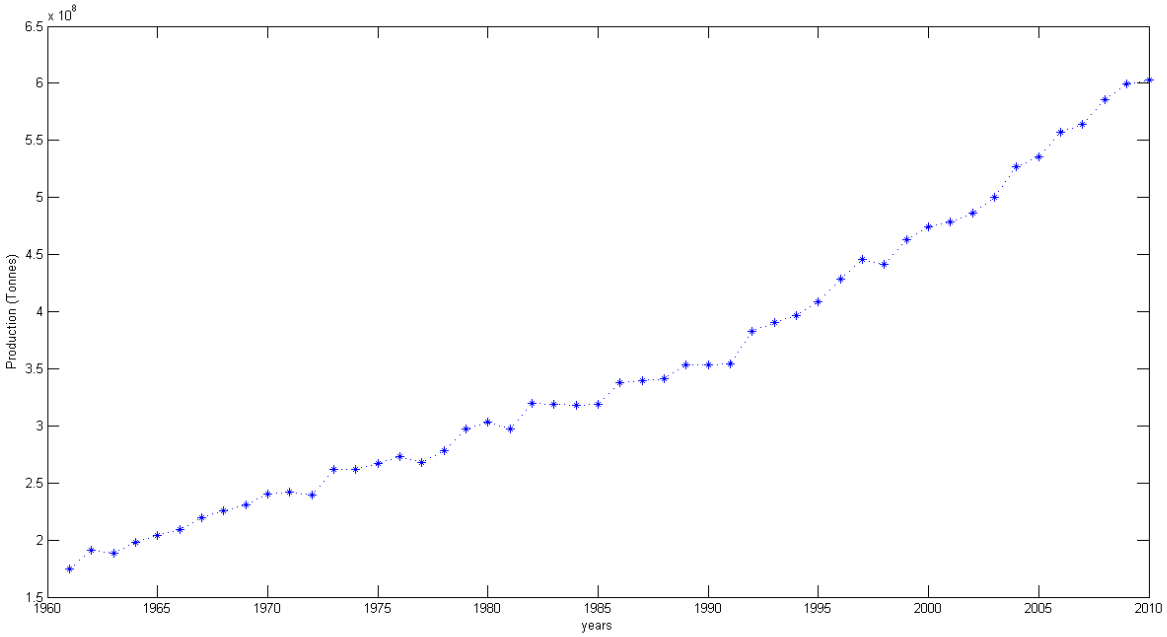


Figure 1 – Global production of fruits [1]

supply chain in fruit handling and processing is depicted in Figure 2. The main stages comprehend the producer, the preparation and packaging central, the wholesaler, the retailer and finally, the consumer. Most of these stages involve in between, storage and transportation of the produce, which are supposed to follow strict guidelines to guarantee that the final consumer acquire the fruit in the best conditions.

There are many factors that can affect fruit quality. These factors go all the way from production to consumption, meaning the whole production line. Temperature, light exposure, rain and water availability on the soil affect the fruit nutritional value. The agricultural practices such as pruning and fertilization affect the production yield, size of the fruit and composition. The ripeness state of the fruit at harvest greatly influences the quality and post-harvest life. The method of harvest may cause more or less physical damage such as bruises, scratches and cuts on the surface that will lead to deterioration by microbial contamination, increased water loss, up-regulation of ethylene production, and a precocious senescence condition. The time fruit take from harvest to cooling under storage conditions, also lead to water loss and alterations in terms of flavour and nutritional quality. Although all these facts are known, there is no way to use them to predict, with precision, fruit quality. However, it is known that 'bad practices' lead to a decrease in quality and more losses along the post-harvest chain. Losses can go from 5% to 25% depending on the product and on the country [2].

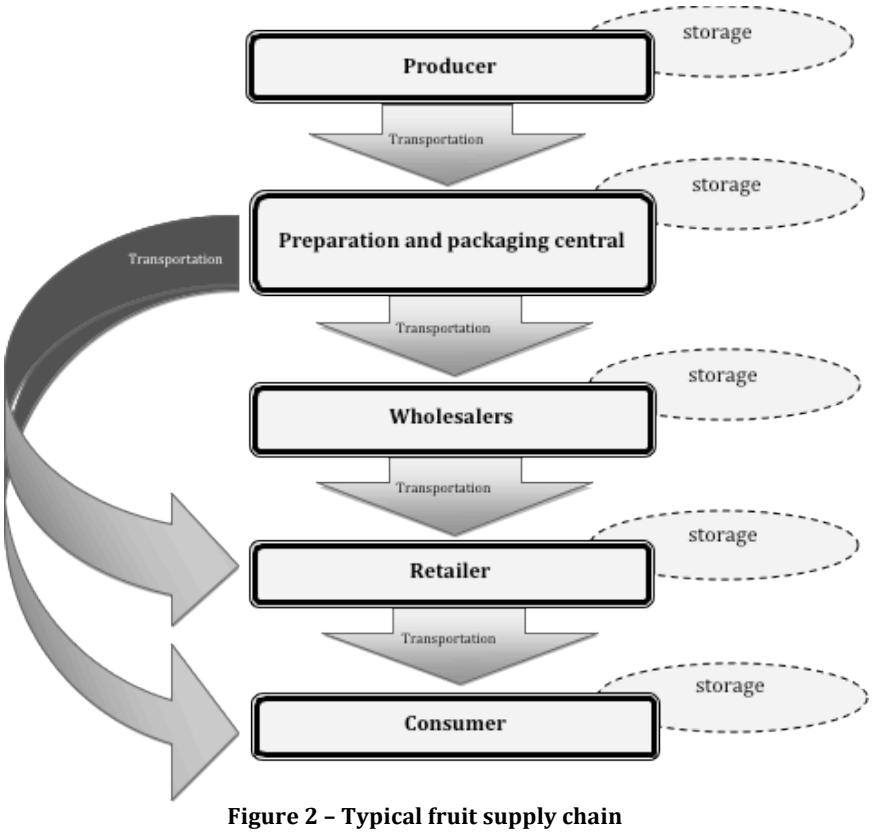


Figure 2 – Typical fruit supply chain

Specific legislation and various systems have been implemented and developed in order to guarantee food safety and quality from the harvest to the consumer. Food safety risks can be biological, chemical and physical. When a problem is detected in a produce or batch, it should be possible to track its entire route back to the respective source. For this to be possible, it is necessary that all stakeholders in the supply chain follow the same standardized systems [3].

European Union regulation EC nº2200/96 for fresh fruit and vegetables presents the norms for quality. This regulation includes the quality categories of size, presentation and the minimum requirements for commercialization. There are three categories: 'Extra Category' (superior quality); 'I Category' (good quality); 'II Category' (commercial quality). The quality norms are specific for each fruit, although they are based in the same basic guidelines that include:

- Product definition;
- Quality definition, such as minimal features and category classification;
- Provisions relating to the calibration, function of the diameter or weight, homogeneity in the calibration;
- Provisions concerning the category and calibration tolerances;

Quality is watched and managed differently for each stakeholder along the supply chain. For producers, quality involves to assure that cultural practices provide a high potential of yield and resistance to crop diseases, as well as uniformity of the produce harvest time. The preparation and packaging central has to contemplate the suitability for processing and preservation. Wholesalers and retailers are interested in freshness and durability. The consumer is also interested in freshness and will buy the produce expecting it to be tasty, with nutritional and health promoting properties. Although, on the consumer side, fruit quality tends to emerge apparently as a subjective concept, the ultimate fact is that, at this stage quality means acceptability, and therefore economical benefits or losses for all the elements involved in the production line.

In Table 1, the most important fruit characteristics in terms of quality are presented. Besides these, presentation and the appearance of the fruit on the shelf, packaging and labels used are known to influence consumers' acceptability.

Table 1 – Quality factors and the characteristics with more interest (Based on Boas Práticas)

Factor	Characteristics of interest
Visual appearance	**Size**: dimensions, weight and volume
	Shape and aspect: irregularity and uniformity
	Colour: intensity and uniformity
	Glow: natural or from wax
	Defects: external or internal morphological, physical or mechanical, physiological, pathological, entomological
Texture	Firmness
	Crisp
	Fibrous
	Hardness
Flavour (smell and taste)	Smell
	Bad smells or taste
	Sweetness
	Acidity
	Astringency
	Bitterness
Nutritional value	Vitamins
	Minerals
	Carbon hydrates
	Proteins
	Fat content
Safety	Natural toxic components
	Contaminants: chemical residues from pesticides and heavy metals or cleaning products
	Mycotoxins
	Microbial contamination

1.2. Internal Quality Parameters of Fruits

Fruit quality parameters may be classified into external and internal. External quality parameters are related with those aspects that may be observed by direct visual inspection or quantified through a straightforward measure. External quality parameters include the size, the shape, the aspect, the colour, the glow and the presence/absence of peel defects. The internal quality parameters, on the other side, are related with those aspects that can only be quantified through extraction of the peel and/or the pulp, resulting in the destruction of the fruit. The internal quality parameters include the texture, flavour, nutritional value and safety (Table 1).

One of the most important components of the marketing standards is precisely based on the internal quality parameters with the objective of increasing fruit economical value. Minimum characteristics of texture, taste, aroma and nutritional values are required.

Fruit selection and sorting provide the market with more uniform batches. This process is usually called calibration. Calibration is a great asset, because it allows increasing the value for higher quality fruit and also because uniform batches increase the perception of quality. Calibration can be performed separating the fruit by size, shape, colour, defects, composition or any combination of the previous.

For many years producers have preferred cultivars that provide fruit with good colour, size and resistance to diseases. Until recently these were the parameters with more importance because they correspond to what consumers see when buying the products. Nowadays there is an increased attention to the internal quality parameters of fruit from all parts, from the producers to the consumers. In some particular cases, such as the protected geographical indication (PGI) and protected designation of origin (PDO) fruit, these internal parameters are quite strict and produce can only be considered under these nominations if these parameters attain certain minimum values.

Internal quality parameters are decisive for consumer final satisfaction. Many times these are perceivable only when the produce is cut and consumed. Consumers base their first choice of fruit in colour and consistency, but when they repurchase, the flavour that was experienced formerly, influences if they buy it or not. A fully satisfactory experience generally leads to a new buy and eventually recommendation of the produce to other consumers.

To evaluate fruit internal quality, specific parameters must be evaluated or measured. The texture is mainly a combination of firmness, consistency and turgidity. Flavour is a very complex property, mainly because of its subjectivity. It can be thought as two separate parts, taste and aroma. The aroma has received increased attention for his role in the flavour quality of fresh produce. The aroma is due to the volatile constituents of the fruit, and these can affect the perception of sweetness and acidity. The constituents of the aroma that influence the flavour can be between 15 and 40 in tomatoes or apples and 3 in bananas. Sweetness, acidity, bitterness, astringency and the relationship among them can be investigated to address taste. Finally, nutritional properties such as vitamins, minerals, carbohydrates, fats and proteins can be evaluated individually, through very time –consuming chemical analysis. As previously indicated, the consumer concerns on the nutritional properties (e.g. antioxidants) and its benefits for health of fresh commodities, namely fruit, has increased dramatically in the last decades and constitutes presently a factor of major importance [4].

The fruit sweetness originates from soluble sugars, particularly fructose, glucose and sucrose. The most common procedures to measure the sugar content use the Brix scale (in Brix degrees; ºBrix). This scale has been developed to measure the sucrose content of an aqueous solution. The juice extracted from the fruit, however, is never a pure solution of sucrose. Besides the other sugars (fructose and glucose), there are other constituents in minor quantities such as acids, proteins and lipids. However, it is a good approximation to assume that the total soluble solids content of juice is essentially determined by its sugars content and hence that the ºBrix reading is a good measure of total sugar content. The ºBrix is slowly becoming part of the standard measures for calibration of fruit.

Advances in technology and algorithms allowed for the implementation of methods to measure or estimate the automatically the ºBrix. These methods can be destructive or non-destructive, or described as invasive or non-invasive, respectively. However, these are not equivalent terms. For example, an electronic nose is truly non-invasive; but methods based on light spectroscopy are indeed invasive, since light has to penetrate the fruit. However, both are non-destructive. In the context of this work the description non-destructive is more appropriate and hence we will adopt it.

The destructive methods are used extensively in the industry. For instance, refractometry is used to measure the ºBrix. Depending on the fruit, part of the peel of the fruit is removed; some of the fruit juice is then squeezed into a refractometer to make the measurement. This method has to be done manually and it is time-consuming. Since the method is destructive and takes some time, the calibration is done using only a few fruit from each batch. Improving the batch evaluation would require many workers in the calibration lines. The disadvantages are obvious in terms of hand labour cost and fruit loss. Hence, the usual procedure is to use a statistically under representative subset of the batch, possibly conveying a false description of the fruit quality [5].

In order to overcome the limitation posed by destructive methods, many non-destructive methods are being developed to evaluate fruit quality. Some examples are based in near infrared spectroscopy, image processing and electronic noses. These techniques allow for real time prediction of internal quality parameters, making it possible to evaluate the quality of each fruit of a large batch. Non-destructive methods have improved significantly in the last years. Having achieved small errors, their use is being steadily adopted in the industry. However these methods are still expensive and the compromise cost/benefit is not always achieved, especially for small/medium size companies.

Until non-destructive methods are more affordable and a broad knowledge about it exists in the industry, calibration will still be done manually trough the destructive methods.

1.3. Aim of the Work

This work is an important component of a bigger project. The project upraised from a partnership between the companies Calibrafruta, MCM-Electronics and the R&D Center of Electronics Optoelectronics and Telecommunications (CEOT, Universidade do Algarve). Calibrafruta is a company that produces automatic grading and sorting lines for fruit calibration. MCM-Electronics, a partner of Calibrafruta, is responsible for the integration of sensors on the lines and for the development of the software that controls the calibrator.

The project aims to develop a prototype for ºBrix determination and its incorporation on the automatic calibration lines. Earlier in the project it was determined that the most promising method would be NIR spectroscopy (in the wavelength range of 700 - 2500 nm). However, since NIR spectrometers in this range are quite expensive, its integration on many calibration lines would not be feasible. So, it was decided to test the visible/near infrared range instead (Vis/NIR), with a spectrometer working in the range 500-1100nm. After the preliminary tests, the results showed feasibility of the Vis/NIR technique and the project proceeded.

This work will address the use of reflectance spectroscopy, specifically in the Vis/NIR region from 500 to 1100 nm, to perform on-line determination of fruit ºBrix in automated calibration lines. The wavelength limits are determined by the spectrometer available on the laboratory (in turn determined by the responsivity of the silicon used in the detector, which has a cutoff at 1100 nm). Further details about the hardware are described ahead.

As described on the following sections, the use of spectroscopic data to make predictions involves a series of steps. It is a difficult task that has to be coupled with chemometrics to obtain useful results. This work aims to develop a test and development platform for the prediction models. Specifically, this platform should be able to perform the following tasks:

- acquire the spectrometer data (spectra) to a computer;
- use the data to calibrate models for the prediction of ºBrix;
- validate the generated models and present quantifiers for the performance;
- test different parameters of acquisition and their effect on overall system performance.

The time of integration of the spectrometer and trigger mode are examples of acquisition parameters. The models will be generated using the Partial Least Squares Regression (PLSR) algorithm. The parameters should be controlled through user interfaces, allowing versatility of operation. Also, real time visualisation of the measured data/results should allow for iterative improvement of the configuration settings.

Software will be developed to address these objectives. The software will be developed in Laboratory Virtual Instrumentation Engineering Workbench (LabVIEW™). LabVIEW™ is a graphical language development platform crated by National Instruments.

Calibrafruta provided the spectrometer, the light sources, the optical components (lenses, fibres) and a small calibration line to simulate the real environment operation.

MCM-Electronics provided a computer with LabVIEW™ to develop the software.

2. State of the art

2.1. Measurements methods for quality evaluation

Figure 3 shows the distribution of results of an Internet search on the keywords "nondestructive" or "noninvasive" in the context of "food" on the different sensor techniques of non-destructive food analysis. It is clear from Figure 3 that vision techniques are more used than any other, that acoustic methods have a major role and that near infrared (NIR) is more used dominates amongst the spectroscopic techniques.

Humans use all their senses to evaluate fruits. So despite the devices used to make measurements or the methods used for predictions, all the available techniques should be integrated in a coherent picture to give an evaluation of the fruit quality as close as possible to the human sensory response [6].

Some of the methods and instrumentation technologies to make this possible are presented below.

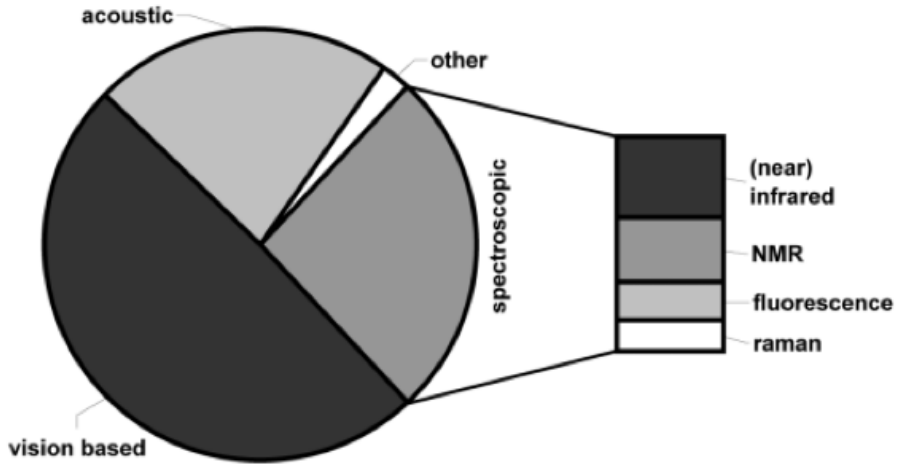


Figure 3 - Distribution of results for an Internet search of the terms "nondestructive methods" or "noninvasive" [6]

2.1.1. Mechanical and Acoustic

Mechanical methods have been used mostly for prediction of the fruit firmness. However some published reports have shown that these techniques can also be used for detection of internal defects, ºBrix and mass prediction [7]. Mechanical methods include impact, quasi-static force, sonic and ultrasonic propagation. Impact technique consists in using a rod or pendulum to impact the fruit. The impact causes the fruit to vibrate and a piezoelectric element, microphones or accelerometers can record this effect. The vibration can then be analysed, either using the duration, peak intensity, peak frequency or bandwidth.

Compression and puncture are considered quasi-static techniques. These consist on applying a force to the fruit and observing the ratio of contraction and relaxation of the tissues. The time and maximum force applied must be well adapted to the species of fruit in test, since it can result in rupture or damage of the tissues. Sonic and ultrasonic techniques use an emitter to project sound waves against the product of interest; the reflected or transmitted waves provide information about the interior of fruits [8–10].

2.1.2. Electric and Electrochemical

Electric technology can be used to evaluate internal properties of fruits. These methods are simple when compared to others, in terms of equipment and data processing needed. The technology is based on the use of sensors for capacitance, inductance, impedance or a combination of these. The data captured by the sensors can be correlated to the dielectric properties of the fruit being measured. The implementation of the method presenting more advantages employs a frequency scan. At lower frequencies (10MHz) the dielectric properties correlate to the external surface and at higher frequencies (1.8GHz) the dielectric properties correlate to the internal tissue [8]. Density is a property that correlates to the internal quality of watermelons. Kato [11] investigated this subject and proposed a method to estimate density using electric capacitance for volume and an electronic balance for mass.

Electrochemical technologies have a promising future in the detection of volatiles. The name ‘electronic nose’ or ‘e-nose’ is often used to express the sensors that try to simulate the human nose. These technologies can be used to detect a specific compound or a group of compounds. It has been stated that the presence of one compound is not as important as the combination of many. Thus this type of sensors should detect a chemical fingerprint that distinguishes the compounds involved [12]. Results show that with the right combination of polymers and semiconductors, specific groups of volatiles can be detected. This can be used to evaluate the pleasing aroma of a fruit, its maturity state and even the presence of some disorders [9]. Despite all the studies in this type of technologies in the last decade, the technologies are not mature enough to reach the industry. The sensors still have a lack of sensibility, lack of linearity and are highly dependent on the surrounding environment. Furthermore, electrical inductance measurements need direct and perfect contact with the fruit, making its implementation on automated fruit grading lines virtually impossible. Che Harun et al. [13] presents a new architecture for e-noses. Future advances in nanotechnology are expected to greatly improve this field.

2.1.3. X-Ray

X-ray has a deep penetration. The energy transmitted through the fruit being evaluated depends on the absorption coefficient, density, thickness and the source strength. Thus the technology can be used to evaluate maturity and internal tissue of fruits. It has also been used to detect the presence of insects. There are various X-ray techniques and the technology is used for many years in medicine. The most common is X-ray radiography that projects a two-dimensional (2D) image of all the layers of the fruit being evaluated. Computed tomography (CT) on the other hand, produces a 3D image making 'slices' of the fruit. CT gives more information about the fruit but is also more complex in acquisition and data treatment. Although this technique is mostly used in laboratory it has been used in commercial applications [8], [9].

2.1.4. Magnetic Resonance

Magnetic resonance (MR) is mostly used in medical applications, but this technique can have many uses in fruit calibration. There are two major fields using this technique, nuclear magnetic resonance spectroscopy (NMR) and magnetic resonance imaging (MRI). Both techniques are based on the same principle. When an atom with non zero spin is immersed in an external static magnetic field, its spin will tend to align with it. However, due to the particle's quantum nature, only discrete levels are allowed, corresponding to a set of discrete directions for spin orientation. The energy differences between these states match the energy of a photon with the so-called Larmor frequency (which is the classical frequency of precession of a magnetic moment around the direction of an external magnetic field). Therefore, irradiating these atoms with electromagnetic radiation at the Larmor frequency (which is typically in the band of radio waves) induces resonant absorption and the promotion of atoms to the highest energy state (which is anti-parallel to the magnetic field). Different atoms have different resonant frequencies. Furthermore, the interaction between atoms and electrons in a molecule induces frequency shifts in the absorption frequencies that allow precise atom identification.

Nuclear magnetic resonance can be used to get information from a specific region of the fruit, where the result is a plot of frequency versus intensity. As for magnetic resonance imaging, the information returned is spatial, giving evidence about the tissues of the fruit. The images may be in 2D or 3D. The former gives less information, but it has been used to quantify sugar content and organic acids, whereas the latter has more application in detecting internal defects, since a structure of the internal tissues can be observed. MR is an expensive

technology and it is difficult to operate with it. There still exist many challenges to make it fast and affordable enough to see its use in on-line calibration [9], [14].

2.1.5. Fluorescence

Fluorescence imaging has been used to detect internal defects on fruits. This technique consists in the excitation of molecules by high-energy light, usually deep blue or ultraviolet light. After the excitation the molecules, usually chlorophyll, respond emitting light in the Vis/NIR band. Fluorescence was also used to monitor stress levels and maturity of fruits [15]. Fluorescence imaging as been implemented in on-line systems for fruit grading. It consists basically of an illumination system emitting the excitation light and a camera with a optical passband filter tuned to the fluorescence emission. Fluorescence imaging systems have been able to detect biological contaminants in fruit [9], [16], [17].

2.1.6. Computer Vision

Fruit quality is often related with its appearance and colour plays a major role in the overall perception of quality by the consumer. Computer vision can be used to classify and quantify fruit parameters. Using a regular charged coupled device (CCD) camera and appropriate algorithms, external parameters can be assessed. Size, shape, surface texture, surface colour and defects are some of these parameters. The main components of a computer vision system are lamps, camera, computer hardware and software. As in human vision these systems depend very much on illumination, particularly in what concerns output stability, uniform distribution of radiance and the shape of the light spectrum. Image processing and image analysis are the core of computer vision. Image processing consists of a collection of methods intended to enhance the quality of the image and image analysis addresses detection of regions of interests (ROI) and its qualification and quantization. The choice of the camera depends on the application. For instance, a grey camera can be used to quantify the surface colour of an orange (mono coloured fruit) but for an apple a colour camera has to be used. Computer vision allows for a major improvement in on-line systems for calibration, speeding up and improving the process of calibration. Improvements in algorithms over the last years allow the classification and distinction of defects like bruises, wounds, bitter pit, frost damage and others [18], [19]. Visible (Vis) spectrum comprises the range between 400 and 780nm. Colour and grey scale images are inappropriate for detecting internal quality parameters because most of the light absorption bands with interest for internal quality classification lie outside the visible spectrum; furthermore, the image recorded by the

cameras is dominated by the light reflected by the fruit surface. Hence, the information available from the cameras comes mainly from the skin.

2.1.7. Near Infrared Spectroscopy

Near infrared spectroscopy (NIR) covers the range from 780 to 2500nm. Many chemical compounds and molecules absorb light within range, making it a focus of great interest. There are other regions of interest further into the infrared (IR), and techniques such as Fourier Transform Infra Red (FT-IR) spectroscopy can go as far as 25 um. However most of these techniques are unsuitable for on-line calibration because of the time needed to perform the measurements.

Figure 4 shows three possible setups for acquisition of NIR spectrum. In the three methods a source emits NIR radiation against the fruit, the radiation penetrates the tissues and its characteristics change due to absorption and scattering. The scattering and absorption depends on the structure, microstructure and chemical composition of the fruit, which means that information can be obtained from the received light. The reflected/transmitted light is collected by a spectrometer, which measures its spectrum and converts it to numerical data. Using many samples and the corresponding spectral data, advanced multivariate statistical techniques can be used to extract useful information.

In reflectance (Figure 4a) and interactance (Figure 4c) spectroscopy modes, the

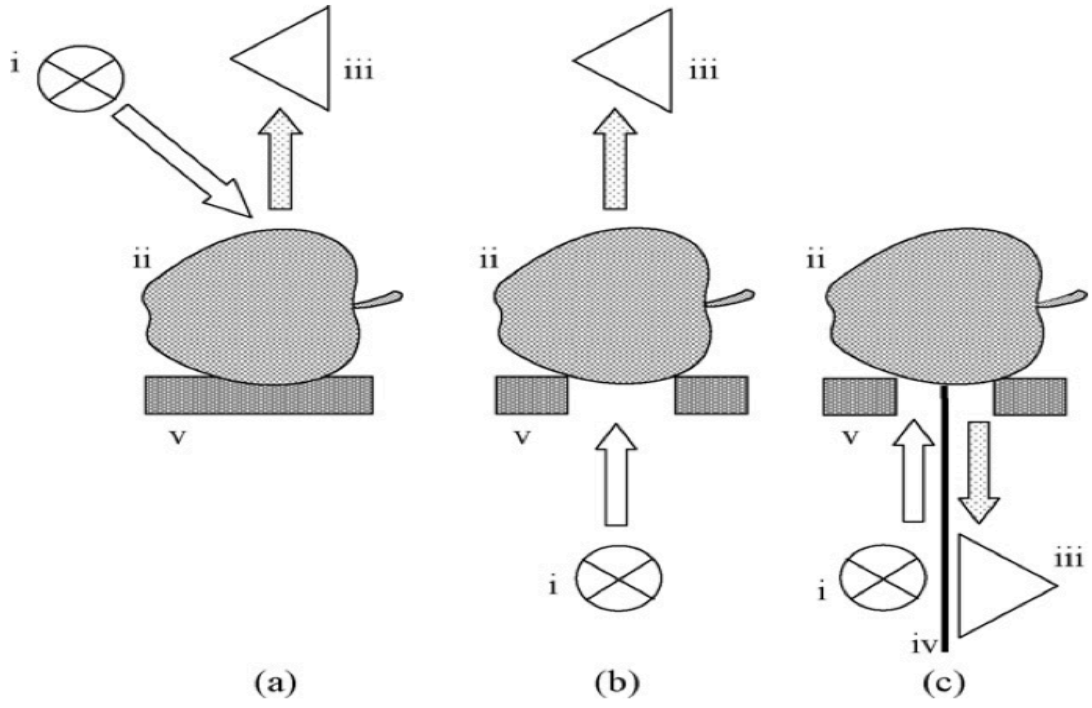


Figure 4 – Setup for the acquisition of (a) reflectance, (b) transmittance, and (c) interactance spectra, with (i) the light source, (ii) fruit, (iii) monochromator/detector, (iv) light barrier, and (v) support. In interactance mode, light due to specular reflection is physically prevented from entering the monochromator by means of a light barrier [20].

radiation is emitted, penetrates slightly into the fruit and some is reflected back. As for the transmittance mode (Figure 4b), the emitted radiation goes through the fruit and the transmitted light is received on the opposite site. Transmittance is mostly used for non-solid materials, because the source needs to be very strong to cross solid matter.

NIR spectroscopy is used in agricultural applications since 1964 [21–24]. Since then, it has gained more attention while more uses have been found for it and technological advances made available simple and economical spectrometers.

NIR spectroscopy has been used for the measurement of moisture, protein, fat, dry matter, soluble solids and water content. Newest developments include stiffness and internal damage. Among the new developments, multi- and hyperspectral imaging and time-resolved spectroscopy are expected to provide new methods of calibration to the industry.

Multi- and hyperspectral imaging provide spatial and spectral information at the same time. Multi-spectral imaging is usually obtained with a set of filters in front of the camera or by using monochromatic light sources [25].

Taking a photo of the same object with different filters provides multi-wavelength information about each pixel of the image. Hyperspectral imagers work on a different principle, scanning a line of the object at a time. For each pixel of the observed line a spectrum is obtained by sending the light through a transmission grating and recording all the spectra (from all the pixels in the observed line) in a CCD. The information from a single line is then stored as a matrix where the column number corresponds to the x-position of the pixel and the row number corresponds to the wavelength. Scanning through several lines produces a hypercube of data where the third dimension is the y-position of the line.

A practical criterion to distinguish between multi- and hyperspectral imaging is based on the number of wavebands. Multispectral imaging acquires few wavebands (generally less than ten) with bandwidths of 5 to 50 nm. Hyperspectral imaging uses tens or hundreds of images at close wavelengths or specific wavebands of interest [9].

Because of the ability to provide spatial information, Multi- and hyperspectral imaging allow to find the location of internal defects. Hence, it is expected to have much more applications in the coming years. However, the process of acquiring this type of images is slow and its application in on-line systems requires the use of a limited number of wavelengths [26], [27].

Time-resolved spectroscopy is a technique where a very short pulse of light is injected into the fruit and collected some distance away. The pulse spreads during its propagation in the biological tissue due to light scattering effects. The transmitted pulse is detected by a fast

detector (a photo-multiplier tube or an avalanche photodiode, for example) and the shape of the spread pulse is measured. Information about the scattering and absorption properties of the tissue may be determined from the shape of the spread pulse (usually in the form of the reduced scattering coefficient and the absorption coefficient, both in units m-1) [20].

2.2. Statistical Analysis (Chemometrics)

Relating instrumental data to quality parameters is a difficult task. Figure 5 shows typical NIR reflectance for some specimens of fruits. Despite being originated from different fruits, the spectra are similar; and spectra from the same species present even more similarities. The similarity is the reason why advanced multivariate statistical methods are needed to extract useful information of the data.

In general there is no obvious or strong correlation between the quantity to be determined (for example the Brix) and any of the spectrum wavelengths. For example, it is common to find two fruits with similar spectra and dissimilar Brix values, but the opposite is also possible. The main problem is that the Vis/NIR spectroscopy has low selectivity and the signal from the constituents of interest is contaminated in unknown amounts by other (mainly) unknown constituents. Correlations between dependent and independent variables in a typical measurement on a physical system should be well above 0.9 if a meaningful model is to be obtained from the measures. When the samples are biological, the concept of biological variability means that the measurements vary from sample to sample due to uncontrollable factors. Correlations of the order of 0.6 to 0.8 are considered as a good starting point for a prediction model. In order to do that, statistical multivariate analysis must be

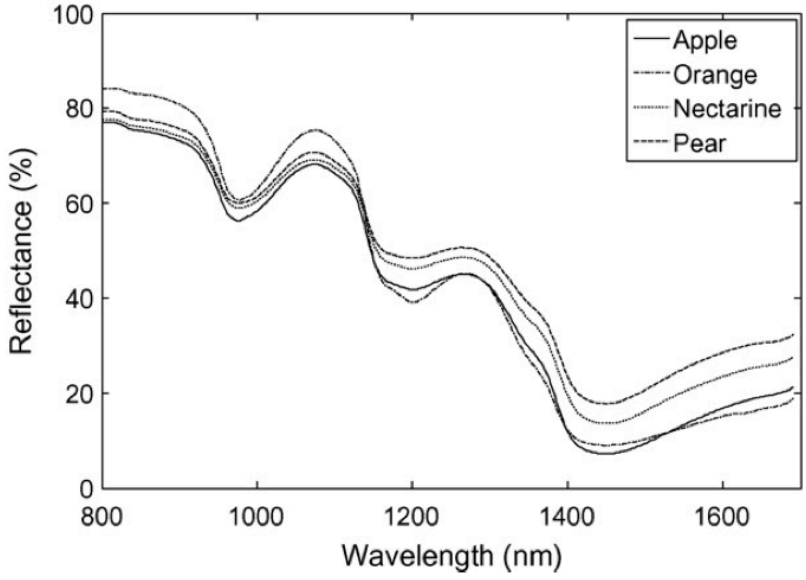


Figure 5 - Typical NIR reflectance spectra of some fruit. The NIR reflectance spectra were recorded using a Corona 45 Vis/NIR diode array spectrophotometer (Carl Zeiss Jena Gmbh, Jena, Germany) [28]

employed, to uncover hidden relations and avoid parasitic influences on data.

Great advances in the use of classical statistics made this difficult task possible. Statistical methods are the basis for data reduction, regression techniques, pattern recognition, multivariate modelling and classification [9]. The advanced multivariate statistical methods can find intricate relationships between the spectrum and the parameters of interest in the fruit. The use of multivariate statistical techniques for these purposes is usually called chemometrics. An overview of the chemometrics techniques is presented below.

2.2.1. Basic Convention

Consider a set of measurements performed on n samples. For each sample a spectrum is obtained. The number of spectrum variables is m (for example, m wavelengths). The full data set is presented in a matrix *nxm*. Hence each row represent a sample and there are n rows; each column represents a variable and there are m columns.

The mathematical operations on the data matrix described below may be of two types: if the operation is performed on each line (that is, "row-wise" or sample-based), we will refer to it as longitudinal transformation. If the operation is performed on each column ("column-wise" or variable-based), we will refer to it as transverse transformation. In general, it is desirable to perform longitudinal methods prior to any transverse method.

2.2.2. Preprocessing

Preprocessing is the general denomination attributed to a mathematical operation applied on the spectrum data before being input to the statistical analysis algorithm. This procedure can be performed on a single step or on multiple steps. Preprocessing is used to linearize the data and to remove extraneous sources of variation. Basic transverse preprocessing procedures are mean-centring (subtraction of the data by the mean, performed for each variable on the data) and variance scaling (also performed for each variable, it consists in the division of the data by the standard deviation). The two steps combined are called auto-scaling. The meaning of auto-scaling is more clearly understood by observing Figure 6: after auto-scaling all variables are centred around zero and have standard deviation equal to 1.

Usually auto-scale is recommended as the minimum preprocessing required before applying the statistical algorithms to the data. Usually much better results are obtained after auto-scaling. This is because after auto-scaling all variables are on scales numerically

equivalent, which is particularly important if the initial variables correspond to different physical quantities (like capacitance and optical transmittance, for example), or if there are orders of magnitude of difference between some variables (optical transmittances near 100% and optical transmittances near 1%, for example).

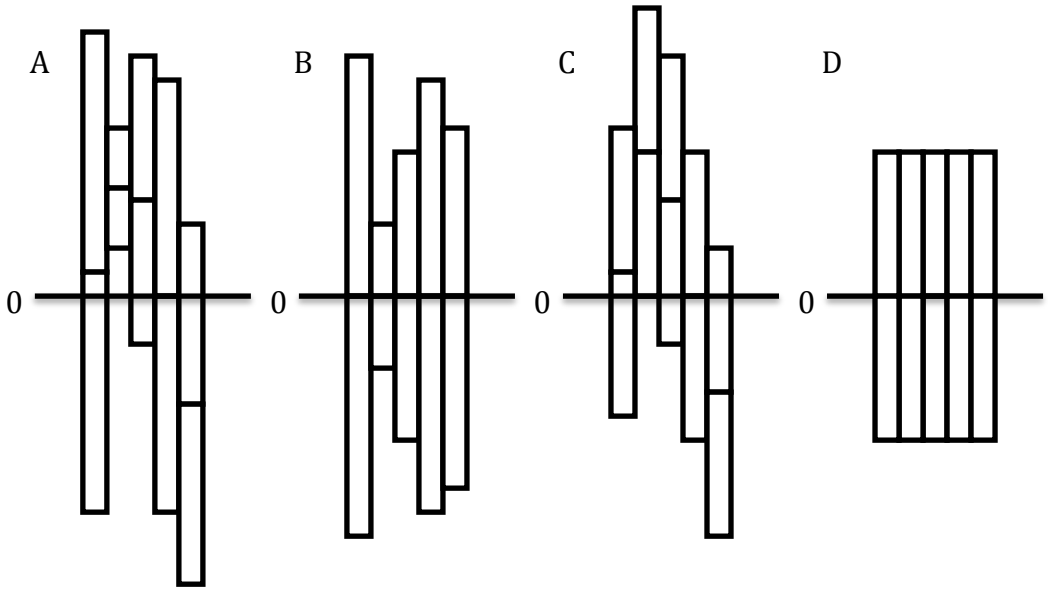


Figure 6 – The data for each variable are represented by a variance bar and its center. (A) Most raw data look like this. (B) The result after mean-centering only. (C) The result after variance-scaling only. (D) The result after mean-centering and variance-scaling. Based on Geladi (1986)[29]

Preprocessing is also important because it may contribute to linearize the data. In turn, this is important because most algorithms expect linear responses of the variables and because linear responses are easier to model. Interfering sources of variation also increase the difficulty of modelling. Instrumental noise and the surrounding environment changes are examples of interfering sources.

Smoothing is a longitudinal transformation that is used when the data is very noisy. It makes use of the neighbouring data to smooth the spectrum. Smoothing is performed on each sample (longitudinal, as stated above) since noise is expected to be independent from sample to sample. Moving averages and Savitzky-Golay [30] are examples of smoothing algorithms.

Following the general rule stated above, longitudinal transforms are performed prior to the transverse ones. Hence, smoothing should precede centring and scaling transformations. Smoothing should be used with caution because it can 'hide' important relationships among adjacent variables.

Scattering effects are also a source of data variability.

Consider the measurement of two samples in the reflectance mode. Suppose that both have the same absorption characteristics but that sample 1 scatters more than sample 2. The

net result is that more light is backscattered from sample 1, resulting in a higher level of reflectance. For biological tissue the scattering properties are essentially constant over a broad range of wavelengths (contrary to the absorption characteristics, that are dominated by the absorption bands). Hence, changes in sample scattering power result approximately in a multiplicative effect on the spectra. This means that the same variable will be scaled differently, depending on the level of scattering. Similar effects can be seen in other measurements caused by physical or chemical effects. The differences caused be these effects can difficult the processing for a statistical analysis. Normalization attempts to correct for these categories of effects by finding characteristics of each sample that should be invariant under scattering-like transformations. Based on the found characteristics all variables are scaled accordingly. Normalization should be exercised carefully in order to obtain good results; otherwise the result may be worse than without normalization. There is a need to distinguish between variance caused by the properties of interest from the intrusive effects. Normalization also equalizes the impact of different variables for a model creation. Simple normalization (1-norm[1], 2-norm[2] or infinite-norm[3]) is a common technique to remove these problems. Other normalization techniques are standard normal variate correction (SNV) and multiplicative signal correction (MSC). SNV performs a weighted norm, giving additional weight to variables that deviate further from the spectrum mean value. MSC is many times preferred because it can account for scaling effects and offset effects. MSC uses a regression method applied to a sample spectrum and a reference spectrum.

Derivation is frequently used to remove offset effects. Derivation of second order is preferred because it can correct for offset and scaling effects (identical to MSC). The Savitzky-Golay algorithm, presented previously, is able to perform smoothing and derivatives at the same time. For these reason the algorithm is very popular, since it is able to smooth the spectrum and to correct the unwanted effects in one single step.

Based on studies of light penetration, scattering and diffuse reflection some transformations have been proposed accounting for the changes when light penetrates tissues. Logarithms and exponentials are some examples of transformations that have been used in spectrum data. Absolute value transformation has also been used after derivation to remove negative values.

---

[1] Normalize to (divide each variable by) the sum of the absolute value of all variables for the given sample.

[2] Normalize to the sum of the squared value of all variables for the given sample.

[3] Normalize to the maximum value observed for all variables for the given sample

2.2.3. Principal Component Analysis

With the amounts of data produced by spectroscopy techniques it is indispensable to use algorithms to find useful information among the data. Principal component analysis (PCA) is a very good tool for exploratory data analysis, information extraction and data compression. PCA can find factors that describe the main tendencies among a set of data. PCA uses a technique for matrix decomposition to make this possible.

In the following subchapters, first it will be explained the PCA principle from the classical approach to the covariance matrix. This has the advantage of having a clear geometrical interpretation. Next, the iterative method will be explained, whose result is the same, but much faster to implement in numerical calculations.

2.2.3.1. PCA explained from the covariance matrix

Consider a matrix of data X with dimensions $n \times m$ where the rows are the samples and the columns are the variables. PCA is based on the determination of the eigenvectors of the covariance matrix of X and the subsequent projection of X on the space spanned by these eigenvectors.

To begin with, the matrix X should be mean-centred, as discussed above. The covariance matrix of X is a matrix $m \times m$ is given by

$$cov(X) = \begin{pmatrix} cov(X(:,1),X(:,1)) & cov(X(:,1),X(:,2)) & \cdots & cov(X(:,1),X(:,m)) \\ cov(X(:,2),X(:,1)) & cov(X(:,2),X(:,2)) & \cdots & cov(X(:,2),X(:,m)) \\ \vdots & \vdots & \vdots & \vdots \\ cov(X(:,m),X(:,1)) & cov(X(:,m),X(:,2)) & \dots & cov(X(:,m),X(:,m)) \end{pmatrix}$$

Equation 1

where the notation $X(:,r)$ represents the column r of X (the symbol : meaning "all rows") and is taken from the standard syntax of MATLAB™. The covariance of two columns (that is, the covariance of two variables) is defined as

$$cov(X(:,r),X(:,s)) = \frac{1}{n-1}\sum_{i=1}^{n} X(i,r)X(i,s)$$

Equation 2

The next step is to find the eigenvectors and eigenvalues of the covariance matrix:

$$cov(X)\vec{v_j} = \lambda_j \vec{v_j}, j = 1, \dots, m$$

Equation 3

There are m eigenvectors v_j and m corresponding eigenvalues l_j. The computation of the eigenvalues and eigenvectors is simple only for low dimensionality. For large values of m the adopted method is usually singular value decomposition (SVD), and it is available in the

libraries for the most used languages. However, SVD is not computationally fast and other iterative methods, such as nonlinear iterative partial least squares (NIPALS), described below, are preferred for intensive computation. For the moment we simply assume that the eigenvectors and eigenvalues are determined by some method.

It can be shown that the eigenvectors of the covariance matrix define orthogonal directions in hyperspace (of *m* dimensions) that correlate better with the data variance contained in *X*. The absolute value of the eigenvalues classifies the relevance of these directions. Hence, to the largest eigenvalue corresponds the eigenvector that determines the principal direction in hyperspace: the one that explains more variance of the data. The second larger eigenvalue corresponds to the second eigenvector, orthogonal to the first, that explains more of the remaining variance of the data. And so on for the next eigenvalues. The principal directions mentioned above are usually known as *principal components*. The series of eigenvectors may be truncated after $h<m$ components have been added, if the eigenvalues become too small. The remaining directions contribute only with minor corrections and may be neglected. This is equivalent to project the initial *m*-dimensional data on a *h*-dimensional space, which allows on one side for easier interpretation of the data and, on the other side, to compress the data.

The last step is thus to project each sample (that is, the vector in the original *m*-dimensional space, corresponding to a sample) onto the new coordinate system defined by the principal components. To do that, the next step is the construction of the matrix of eigenvectors, obtained by juxtaposition of the eigenvectors:

$$W = [\overrightarrow{v_1}\overrightarrow{v_2} \dots \overrightarrow{v_m}]$$

Equation 4

This is a *m x m* matrix and may be used to perform the projection of the original data (*X*) on the new axis, creating the projected data, X^{proj}:

$$\left(X^{proj}\right)' = W'X' \Leftrightarrow X^{proj} = XW$$

Equation 5

The new data matrix X^{proj} has again *n* rows corresponding to the *n* samples, and *m* (or, usually, $h<<m$) columns, corresponding to the *m* (or $h<<m$) principal components. Row *j* of X^{proj} contains the coordinates of the *j*-th sample in the principal components axis.

Figure 7 illustrates the concepts described in this section.

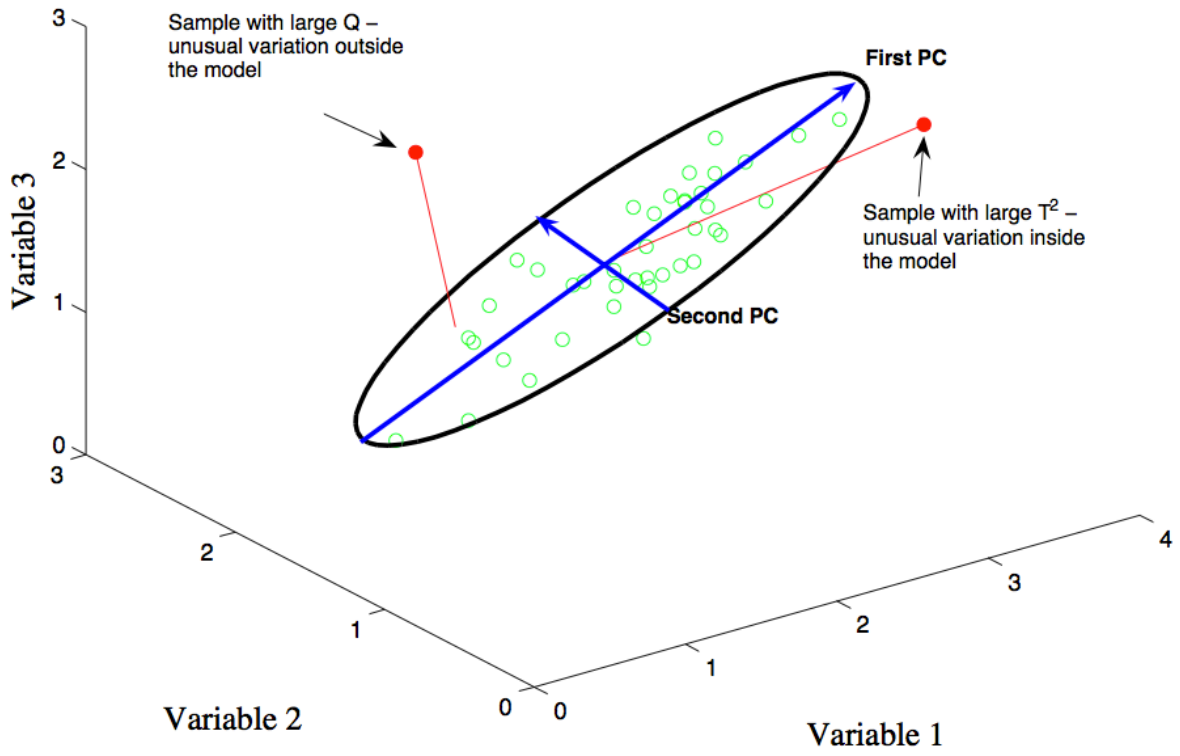


Figure 7 - Graphical representation of Principal Components Analysis [31]

2.2.3.2. PCA explained from NIPALS

As explained above, the SVD is slow for intensive computations and the NIPALS [32] method has been developed to overcome this difficulty. It is a fast iterative method that converges rapidly to the same PCA solution obtained via SVD.

If X is a matrix of rank r, it can be decomposed as a sum of s matrices of rank 1. Additionally the rank 1 matrices, M_h, can be decomposed as outer products of two vectors t_h and p'_h. The t_h are the scores and keep the information of how samples relate to each other. The p'_h are known as loadings and contain information on how the variables relate to each other.

$$X = M_1 + M_2 + \cdots + M_h + \cdots + M_s = t_1 p_1' + t_2 p_2' + \cdots + t_h p_h' + \cdots + t_s p_s'$$

Equation 6

The relation with the description given in terms of the covariance matrix is as follows. Consider each row of X as a vector (this vector corresponds to one sample and it is defined by the values taken for each variable). Then, the scores are the components of that vector along the principal components (corresponds to X^{proj}) and the loadings are the components of the principal components along the original axis (corresponds to W). Equation 6 is thus equivalent to the inverted form of Equation 5, but in a iterative form. Hence, the first term $t_1 p'_1$ corresponds to Equation 5 when only the first component, $p=1$, is retained; the sum of

first and second terms, $t_1p'_1 + t_2p'_2$, corresponds to Equation 5 when the first and second components, $p=1,2$, are retained, and so on.

The t_h and p'_h pairs are ordered by the amount of variance that they describe. Meaning that the first pairs describe more variance on the data than the second pair and so on (exactly as seen in the covariance matrix description). For these reason, after a reasonably small number of h pairs the decomposition can be truncated and the result will have a small amount of error. The calculated h pairs are called principal components or factors.

$$X = t_1p'_1 + t_2p'_2 + \cdots + t_hp'_h + E = TP' + E$$

Equation 7

Equation 7 is presented in a compressed matrix form where T is a matrix of t columns and P' is a matrix of p' rows. E, a matrix containing the residuals is used to represent the error. The decomposition process is based on the eigenvector and eigenvalue theory.

Nonlinear iterative partial least squares (NIPALS) is an algorithm to calculate the principal components based on the principles presented above. NIPALS does not calculate all the principal components at once. It calculates one principal component at a time. E is calculated subtracting X from the principal component.

The NIPALS algorithm as presented on Geladi (1986) [29] is shown below.

1) take a vector x_j from X and call it t_h: $t_h = x_j$
2) calculate p'_h: $p'_h = t'_hX/t'_ht_h$
3) normalize p'_h to length 1: $p'_{h\,new} = p'_{h\,old}/\|p'_{h\,old}\|$
4) calculate t_h: $t_h = Xp_h/p'_hp_h$
5) compare t_h in step 2 with t_h in step 4. If they are the same, stop. If they still differ go to step 2.

These are the steps for one component. If more components are needed, then X is replaced by the residual $E_1=X-t_1p'_1$ and, in general, the iteration for the n-th component starts from the residual $E_{n-1}=X-t_{n-1}p'_{n-1}$.

Table 2 is presented below for further comprehension of what was written above. It will also be useful for the next sections. The table includes already the notation for Y blocks. Until now we had only the X blocks, since PCA only applies to X blocks. However, we will be interested in Y blocks in the next sections, and the table includes them already for the sake of completeness.

In connection with the first reference to the *Y* block just made, it is also important to understand what is the meaning of calling independent block to *X* and dependent block to *Y*. The independent block or X_{nxm} matrix contains the measured values by the instruments. The dependent block or Y_{nxp} matrix contains the values of interest, those that one must predicted from the *X* values and, in that sense, depend on *X*.

Usually what is intended is to model a behaviour creating a relationship between *X* and *Y*. A model can then be represented by $Y = f(X)$ where $f()$ is a function (model) that can transform *X* into *Y*. In an academic study both *X* and *Y* may be easily measurable and the

Table 2 - Relevant symbols and their meaning, based on Geladi (1986) [29]

Symbols	Meaning
$\|\ \|$	Euclidian norm
i	a dummy index for counting samples (objects)
j	a dummy index for counting independent (x) variables
k	a dummy index for counting dependent (y) variables
h	a dummy index for counting components or factors
n	the number of samples in the calibration (training) set
m	the number of independent (x) variables
p	the number of dependent (y) variables
a	the number of factors (or components) used (<rank of x)
r	the number of samples in a prediction (test) set
x	a column vector of features for the independent variables (size $m \times 1$)
y	a column vector of features for the dependent variables (size $p \times 1$)
X	a matrix of features for the independent variables (size $n \times m$)
Y	a matrix of features for the dependent variables (size $n \times p$)
b	a column vector of sensitivities for the MLR method (size $m \times 1$)
B	a matrix of sensitivities for the MLR method (size $m \times p$)
t_h	a column vector of scores for the *X* block, factor *h* (size $n \times 1$)
p'_h	a row vector of loadings for the *X* block, factor *h* (size $1 \times m$)
w'_h	a row vector of weights for the *X* block, factor *h* (size $1 \times m$)
T	the matrix of *X* scores (size $n \times a$)
P'	the matrix of *X* loadings (size $a \times m$)
u_h	a column vector of scores for the *Y* block, factor *h* (size $n \times 1$)
q'_h	a row vector of loadings for the *Y* block, factor *h* (size $1 \times m$)
U	the matrix of *Y* scores (size $n \times a$)
Q'	the matrix of *Y* loadings (size $a \times p$)
M_h	a rank 1 matrix, outer product of $\mathbf{t_h}$ and $\mathbf{p'_h}$ (size $n \times m$)
E_h	the residual of *X* after subtraction of *h* components (size $n \times m$)
F_h	the residual of *Y* after subtraction of *h* components (size $n \times p$)
b_h	the regression coefficient for one PLS component
I_n	the identity matrix (size $n \times n$)
I_m	the identity matrix (size $m \times m$)

motivation to search for the model purely theoretical. In industrial applications, however, the *X* variables are easily measurable but the *Y* variables are not. The motivation to find a model is thus to provide predictions for *Y* from *X*.

Methods for generating this type of models will be presented in the next sections. In this work, *X* is the reflectance spectrum measured by a spectrometer and *Y* is an internal quality parameter of a fruit such as ºBrix. The reflectance spectrum is the easily accessible measurement, since it is performed without contact with the fruit, and the Brix is the hard to perform measurement, since it involves manual destruction of the fruit. The final goal is thus to use the NIR spectrum to predict the ºBrix, without destroying the fruit.

2.2.4. Linear Regression

The aim of a regression is to find a relationship between a set of data *X* and a property of interest described by *Y*. After a good model has been generated there is no need to measure *Y* because it can be estimated, or predicted, using *X*. The predicted internal quality properties are represented by $\hat{Y}$.

Multiple linear regression (MRL) is used to estimate *Y* using a linear combination of the *X* variables (which are spectral response values in our case). Equation 8 is the mathematical representation of MLR. *Y* is estimated using *X* and a combination of linear values *B*. The *B* matrix values are usually called the sensitivities or the regression coefficients. *E* represents the error, similarly to PCA.

$$Y = XB + E$$

Equation 8

The matrix *B* needs to be determined in such a way that minimizes the error (or residual) *E*. The most common way of doing this is using the least squares method shown in the equation below.

$$B = (X'X)^{-1}X'Y$$[4]

Equation 9

There is a potential problem with this approach, because the inverse of *X'X* might not exist. This happens when some variables are collinear (meaning that there are some variable(s) in *X* that have a linear relationship) implying that the *X* matrix is non-invertible. This case is also referred to as zero determinant matrix or singularity.

[4] This equation is easy to obtain if we assume that *Y* and *X* have a perfect linear relation. Then *Y=XB*. Only square matrices are invertible and *X* is not necessarily square. Hence we multiply by *X'* to get a square matrix: *X'Y=X'XB*. Now we apply *(X'X)^(-1)* to both terms and get *(X'X)^(-1)X'Y=B*. This is not a proof, but the same solution is derived from a minimization procedure when *E* is not zero.

Exact collinearity is rare for real data since there is always some random noise in the measurements. However, near-collinearity is very common in spectroscopic measurements, since the spectrum in two neighbouring wavelengths is very similar. We will refer to near-collinearity as "high collinearity", following the nomenclature of the main bibliographical references. In the presence of high collinearity the regression coefficients can be calculated but the result will be very unstable. Small variations in the new data (noise) will produce great changes in the results. High collinearity also increases the probability of overfitting the model. Overfitting means that the model fits well the known data but will not work well for new data that has not been used in the model.

Even if a matrix does not have collinearity, the number of samples (n) in X must always be greater or equal to the number of variables (m). Usually the number of variables in a spectrum far exceeds the number of samples taken. This can be solved removing some of the variables, but this process requires a great knowledge about the samples and measurements being taken. Some undesired (noisy) or uninteresting (in a band that is not useful) wavelengths can be removed. Due to all the problems stated usually MLR does not perform very well.

Principal components regression (PCR) solves some of the problems in MLR. In fact PCR is a two-step procedure, a combination of the methods used on PCA and MLR.

The first step is the PCA. The X matrix of data is decomposed into matrices of scores and loadings. In the second step a regression will be made. The scores matrix T can then take place of X in Equation 9. Since the scores are orthogonal the matrix inversion gives no problem.

$$B = (T'T)^{-1}T'Y$$

Equation 10

Care must be exercised in the choice of the number of principal components used in PCR. Too many components can lead to overfitting and to few will lead to poor results. Small principal components can be removed to avoid collinearity and eliminate some noise. Because the first factors retain more information usually a few are sufficient.

The use of PCR solves the problem of MLR but another problem is generated. While PCA guaranties that the variables with more variance are represented in the first principal components, these variables might not contain any relevant information to predict Y. The removed variables may contain relevant information, while the used ones may contain only ‘noise’.

Partial least squares (PLS) was presented by Wold to solve the problem of PCR. PLS has a way of selecting the most relevant variables to predict *Y*.

Since partial least squares regression (PLSR) is in the base of this work it will be explained in detail in a section ahead.

2.2.5. Nonlinear Regression

Some data may present a nonlinear behaviour. In these cases a nonlinear regression technique may be more suitable. Nonlinearities can be detected when observing plots from the errors of predictions (predicted values minus measured values) versus the predicted values. A curvilinear trend suggests the data may contain nonlinearities. When linear regression methods are used in this type of data sets, nonlinearities are often considered noise and therefore important information could be discarded.

Artificial neural networks (ANN) and kernel-based methods are examples of nonlinear techniques. Nonlinear techniques normally apply linear methods, such as PCA or PLS, for data reduction purposes.

Artificial neural networks have their name due to the concept of neuron. Neurons in this case, are computational methods capable of calculating the weighted sums of its inputs and apply a nonlinear function to calculate the output. Generally the most applied ANN is called a multilayer perceptron (MLP). A multilayer perceptron usually consists of a three-layer network, but the network can be far more complex. In a three-layer network there is an input, a hidden and an output layer. The layers are connected, input to hidden and hidden to output. All layers are composed of many neurons, each neuron from one layer connects to all neurons of the next layer. An ANN learns by changing its input weights and the threshold for outputting a result. In supervised training, this may be achieved by feeding the ANN with a collection of known cases and updating weights and thresholds according to a specified training rule. This general procedure may be applied, in particular, to find the best (nonlinear) function that models the relation between the *X* and *Y* blocks [20], [33], [34].

Kernel-based techniques are becoming more popular because they are easier to understand and the results simpler to interpret, contrary to ANN. These techniques extend the data to a space of nonlinearity, a feature space (a feature space is an abstract space where each feature of the sample is represented as a point. The dimensions of this space depend on the number of features used to describe the samples). In this space, kernel functions define measures of similarity between the samples spectrum. The kernel functions are diverse. Two

of the most common are polynomial and Gaussian functions. Support vector machines (SVM) also belong to the kernel-based techniques [20], [35].

Despite seeming really promising, nonlinear techniques have not provided superior results when compared to the linear ones when applied to the field of NIR spectroscopy [20], [36].

Automatic (on-line) calibration of internal quality is based on the use of models. The models can find intricate relationships among the data collected by different instrumentation techniques. This allows for classification and quantification of fruit parameters without damaging the fruits.

There are many factors that affect the capacity of the models to predict and classify correctly. A list of the factors that are required to make a good prediction model are presented below [4]. Obviously, these orientations should be observed independently of the statistical method employed.

- Measurements should be made with precision;
- A good correlation between the measurement and the parameter to be predicted must exist;
- The correlation must hold for different parcels and campaigns;
- The measurements should be made using fruits from many different trees, parcels and campaigns;

2.2.6. Model Accuracy and Validation Quantifiers

The regression techniques presented above are the first step in chemometrics. Usually this step is called calibration or training. The samples used to build the model are known as the calibration samples and constitute the calibration set. The regression algorithms find the relationships between X and Y, but there is still a need to check the validity of the model. So, an extra step must be considered to test it. This step is always necessary to insure that the model will provide results within a small margin of error when applied to new data, and it is generally known as "validation".

Validation is performed on two steps. The first step is called cross-validation (CV) and involves only the samples used for calibration. The second step is called external validation (EV) and is performed on samples totally independent of the calibration set.

In CV the calibration data set is split successively in combinations of two subsets. At each splitting the larger subset is used to create a model and the smaller subset is used to perform the validation of that model. The process is repeated for different combinations of

data splitting in such a way that all samples are used for validation at least once. In the so-called leave-one-out scheme, CV is performed leaving only one sample out at a time. However, it is usually more adequate to leave out a group of samples. The fundamental idea behind CV is to insure that the models created at each data splitting do not have 'knowledge' about the samples used in validation. On the other side, it also insures that all the samples are used in both steps of calibration and validation. The objective of CV is to quantify the error of the model $Y = f(X)$. The global error delivered by CV is calculated from the mean of the errors obtained at each data splitting. More detail on the quantification of the errors is given below.

The results from CV must be interpreted with caution. Since the calibration data usually comes from the same batch, the model is attuned for that samples and the results will have an artificially small margin of error. This is caused by great similarities between samples. Therefore, the results of CV can be deceiving because the model will perform fairly worse when used with new data. One to way to circumvent this problem is to start from a calibration set containing samples from different batches. The more diverse and independent the calibration samples, the better will be the model performance.

Independently on how well CV has been done, it is always recommendable to perform an external validation (EV). EV consists in using new and independent samples to test the model. These samples, constituting the external validation set, should chosen from batches not included in the calibration set. Hence, one insures that the new samples do not have any relation at all with the calibration set. EV usually gives worst results than CV. This is understandable since the model does not have any 'knowledge' about the new data. Therefore, the EV results are generally more representative of the future predictions than those of CV. However, when a good model has been built from a very broad and representative calibration set, it is expected that the errors obtained through CV and EV are similar. This is one of the fundamental characteristics to be expected from a good prediction model.

Once a model has passed the stages of CV and EV it is ready for application in real life, where it is used for prediction purposes only. However, confronting again the predictions against the true values means effectively a new EV. EV's should be performed routinely as model validity checks.

The usual parameters to evaluate the model performance are described next. The main quantifier to evaluate model performance is the root mean squared error (RMSE). The equation is

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_p}(\hat{y}_i - y_i)^2}{n_p}}$$

Equation 11

where n_p is the number of validated samples, and $\hat{y}_i$ and y_i are respectively the predicted and measured values.

RMSE can be distinguished depending on how the validation was performed. Root mean squared error for cross validation (RMSECV) or root mean squared error of prediction (RMSEP) for quantifying the error in EV. This measure calculates the average uncertainty expected for future predictions. A way of interpreting this measure is that approximately two thirds of the predictions will have an error inferior to the RMSE value. It is also common to calculate RMSEC, the root mean squared error for calibration. This is obtained by using all the calibration samples to create the model and then use the same samples to perform the validation. This is also known as internal validation. RMSEC is usually smaller than RMSECV and RMSEP, but the three tend to be similar in a good model.

Standard deviation ratio (SDR) is another important quantifier. SDR measures the model prediction ability. SDR is defined as a ratio of the standard deviation of the measured variables over the RMSEP. An SDR between 1.5 and 2 means the model can distinguish between high and low values; between 2 and 2.5, predictions can be performed with rough errors; above 2.5 good predictions; and excellent predictions can be obtained for SDR above 3.

The correlation parameter R, is also very useful. This quantifies how well the predicted and real values relate to each other. A value of 1 means the correlation is perfect and 0 means there is no correlation at all.

The three quantifiers presented are usually found in the literature. These can be found expressed in another way or with other names but essentially they will mean the same. For this work another quantifier was used, the percentage of predictions that have an absolute error below 1 ºBrix (designated in the graphs by %<1). This quantifier is important for acceptability when the models are used in the industry.

2.2.7. Model Robustness

The Accuracy of a prediction model can be greatly affected when challenged by new samples with characteristics very different from those of the samples used in calibration. A model that maintains its effectiveness despite uncontrollable changes by external factors is said to be robust. Common factors that can impact the predictions are temperature effects or

drifts, calibration transfer and samples from different batches. Different robust methods have been created to address these and other problems. This section will present an overview of the problems and some solutions to obtain robust models. It is very important to understand what can cause a 'good' model to fail.

Samples from different batches can have very different behaviours from those used for the calibration ("behaviour" here means the relation between spectral features and internal quality parameters). Examples that can cause these changes in behaviour are sun and water exposure or the availability of nutrients in the soil. Even the position of the fruit within the tree (sunny vs. shady sides) can result in important alterations in the spectra. So, different batches of fruit have in general distinct behaviours and batches with characteristics distant from those used in the calibration usually are not correctly modelled. This problem is very common, since it is difficult to gather samples from many different sources and it is impossible to model all the variance that can exist within a given fruit species. So, the simplest solution for this problem is to perform a calibration step with as many as possible samples, representing the widest possible variety of characteristics. To make the model even more robust, the model should be updated, at least, every year.

Ambient temperature, the temperature of the fruit or even the temperature of the measurement equipment can affect the measurements. Also the hardware tends to change with the wear. Some of the problems are uncontrollable. Some solutions have been presented in the literature, as mentioned in the following. Typically this type of problems results in bias in the predictions. Therefore, a pragmatic approach to improve the robustness is to find the appropriate bias correction in each case. This may be done by recalibration. Recalibration means picking some fruits of a new batch (typically around ten) and infer the bias correction by looking into prediction and true (destructive) values of the internal quality of interest.

Ambient temperature effects can be addressed by controlling the temperature of the environment. Also the temperature of the fruits can be stabilized, although this represents a big challenge. Both problems may be tackled at the same time by building calibration models incorporating samples at different temperatures, in a way that covers the temperature range of interest. The temperature of the environment or the sample can even be inserted into the calibration data. Additionally different models can be created for specific temperatures [37]. Roger et al. (2003) [38] included as a preprocessing technique the external parameter orthogonalisation (EPO) algorithm to remove the temperature bias.

Calibration transfer represents one of the greatest challenges for the industry. The instrumental response of the measurement equipment is different from unit to unit, even

within the same model. This will produce unexpected results when a model was created with samples from one equipment and is used with another one. Another problem is the drift of the equipment after some time of usage. Different instrument standardization techniques have been proposed to address the problems described above. Mapping the results of a master equipment (the one used for creating the model) into the slaves (the new equipments) is one of the possibilities. Greensill et al. (2001) [39] compared a series of methods for calibration transfer using various techniques for transforming the data with good results. Usually these methods use data of a small number of samples from both instruments. The data is then used to create a transform that maps behaviour into another or to create a model that eliminates the differences [20].

This section ends the chemometrics section. For more details on chemometrics, Geladi (2003) [40] presents a review covering the history of chemometrics, exploratory data analysis, classification, curve resolution and multivariate calibration. Geladi (2004) [41] presents examples and should be read to get a deeper understanding of the subject.

2.3. Industrial Applications

Some companies present non-destructive solutions for external and internal quality assessment. Below some of these solutions are presented. Most of the solutions presented are designed for on-line automatic sorting. The optical technology is presently the most used because of its versatility, either in terms of imaging or in terms of spectroscopic solutions (refer again to Figure 3). Some solutions using mechanical methods are also presented [42].

Brimrose (http://www.brimrose.com/) developed the *Luminar 3030 Free Space Process NIR Analyser* that can be placed directly in the calibration line. The device can be used for sorting of apples, pears and oranges. It can perform real-time analyses of sugar content, pH, acidity, firmness and brix. *Luminar 3030 Free Space Process NIR Analyser* uses an Acousto-Optic Tuneable Filter combined with NIR (AOTF-NIR). The company also presents hand held and bench solutions.

Maf Roda Agrobiotic (http://www.maf-roda.com/) presents the *Globalscan*, *Insight NIR* and *Optiscan* all for use in on-line applications. *Globalscan* and *Optiscan* are both artificial vision systems. They can be used for colour, diameter and defect sorting. *Insight NIR* detects internal characteristics using a spectrum analyser. It can be used for colour, brix, dry matter percentage and oil percentage. Maf Roda Agrobiotic solutions are designed for a wide variety of products ranging from apples or pears to potatoes and green vegetables.

Greefa (http://www.greefa.nl/) on-line graders are designed for size, colour and weight sorting. External and internal characteristics/defects can also be sensed using *iQS* and *iFA* systems respectively. Intelligent Quality Sorter (*iQS*) uses a series of cameras and mechanical rollers to address external quality evaluation, it can take up to 70 pictures of one fruit to cover the entire surface. Intelligent Flavour Analyser (*iFA*) can sense 'taste'. It uses a halogen source with a spectrum analyser to predict brix, internal brownness and core rot.

Unitec (http://www.unitec-group.com/) employs Vis/NIR spectrometers for detection of internal quality characteristics. Their offer includes the *QS_300* a portable analyser and *QS_ON LINE* for in line detection of sugar content, consistency and ripeness degree. Unitec also presents the *ULTRAVISION* for optical selection of fruit external defects.

Aweta (http://www.aweta.nl/) offers the *Powervision-3D* for external defects detection and the *Inscan IQA* for internal quality. *IQA* uses NIR spectroscopy combined with chemometrics to evaluate sugar content, maturity, firmness and internal flaws. IQA can be used in a variety of fruits and vegetables. Aweta also has a system with an acoustic firmness sensor (AFS) for internal quality assessment. The sensor measures the products vibration to identify rottenness and freshness.

3. Partial Least Squares Regression

This section will be used to describe partial least squares regression (PLSR) in detail.

As stated previously partial least squares regression overcomes the problem in PCR. PCR finds components that represent a great amount of variance in a data set and MLR tries to establish the best relationship between *X* and *Y*. Partial least squares regression has the best of both, it finds components in *X* that characterize the greatest amount of variance and correlate well to *Y*.

There are several methods of calculating a PLSR model, one of the most common methods is NIPALS [43]. PLS can be thought as containing two outer relations, one inner relation and one mixed relation.

The outer relations are the decomposition of *X* and *Y* matrices into scores and loadings. Equation 7 represents the decompositions of *X (=TP'+E)* and below Equation 12 represents the decomposition of *Y*. For the *Y* block, *U* represents the scores and *Q* the loadings (note that in the following *Y* is, in general, a matrix, with *n* rows/samples and *p* dependent variables/columns).

$$Y = UQ' + F^*$$

Equation 12

The inner relation can be obtained by a regression between the scores *T* and *U*. This would be the PCR method and as stated previously a model built this way doesn't perform quite well. To overcome this problem PLS shares information between each block while they are decomposed. To guarantee that, while sharing information, the blocks remain orthogonal and a new set of variables is added to the algorithm. These variables are the weights *W*.

The PLS algorithm [29] is presented below. It is an algorithm that calculates the scores and loadings iteratively, stopping when a convergence criterion is reached.

The PLS decomposition is started by selecting one column of *Y*, y_j, as the starting estimate for u_1. Usually the column of *Y* with the greatest variance is chosen. Of course, in the case of univariate *y*, $u_1 = y$. The algorithm starts with u_1, meaning that the first component of loads, scores and weights is computed iteratively. After the first component has been calculated, the next component is calculated, starting by u_2, and so on .

1) take $u_{start} = \text{some } y_j$

In the *X* block calculate the estimate for the weight in this step as:

2) $w' = u'X/u'\text{u}$

3) $w'_{new} = w'_{old}/\|w'_{old}\|$ (normalization)

the estimate for the score in this step:

4) $t = Xw/w'w$

In the Y block the loading estimate in this step is:

5) $q' = t'Y/t't$

6) $q'_{new} = q'_{old}/\|q'_{old}\|$ (normalization)

update the score *Y* score to

7) $u = Yq/q'q$

Check convergence:

8) compare t in step 4 with the previous iteration. If they are equal (within a certain rounding error) go to step 9, else go to step 2.

Calculate the X loadings and rescale the scores and weights accordingly:

9) $p' = t'X/t'\mathrm{t}$

10) $p'_{new} = p'_{old}/\|p'_{old}\|$ (normalization)

11) $t_{new} = t_{old}/\|p'_{old}\|$

12) $w'_{new} = w'_{old}/\|p'_{old}\|$

The regression coefficient (*b*) for the inner relation is then calculated:

13) $b = u't/t't$

p', q', w' and b should be saved for prediction; t and u can be saved for diagnostic and/or classification purposes.

The steps above are for one component only. If more components are required there is a need to calculate the residuals.

14) $E_h = E_{h-1} - t_h p'_h$; $X = E_0$

15) $F_h = F_{h-1} - b_h t_h q'_h$; $Y = F_0$

16) if all the components have been calculated stop, else go to step 1 replacing *X* and Y by E_h and F_h respectively.

The mixed relation can be expressed as $Y = TBQ' + F$. It is important to note that the intention is to minimize $||F||$. The inner relation is expressed in B, the regression coefficients matrix. With this algorithm PLS finds the components that represent the major amount of variance in *X* that is important to predict *Y*.

Regression becomes useful when it is used for prediction. The prediction is done decomposing a new *X* block and building $\hat{Y}$. As stated above p', q', w' and b from the calibration part need to be saved for the prediction. The steps presented below are used for the prediction.

To decompose the X block:

1) $\hat{t}_h = E_{h-1} w_h$

2) $E_h = E_{h-1} - \hat{t}_h p'_h$

For building the Y block:

3) $Y = F_h = \sum b_h \hat{t}_h q'_h$

the summation is done along *h* for the desired number of components.

Note that the scores and loadings calculated in PLS are not the same as those calculated in PCA and PCR. They can be thought of, however, as PCA scores and loadings that have been rotated in a manner that makes them more relevant for predicting *y*.

This is incorporated in the nomenclature. The components in PLS are referred to as *latent variables* whereas they are named *principal components* in PCA.

Similarly to PCA it is important to be careful with number of components chosen. Too much and the model will be overfitted, too few and the model will not perform well.

The number of components to use in prediction is a very important property. The maximum number of components that can be calculated is equal to the rank of the **X** calibration matrix. Since nowadays computer performance is not a problem, all the components can be calculated in the calibration and used for validation. After that, the amount of error between the prediction and the real values can be calculated. This can then be combined with a threshold value, or other methods, to decide the number of components to use. It is also common to represent the quantifiers graphically and use this to make a decision. An example is presented below.

For this example, 680 spectra samples were used. The spectra were obtained from 'Rocha' pears in our setup, described in the next section. For cross-validation 136 samples were left out of the calibration, representing 1/5 of the total number of samples.

The results of the model validation for the data set are presented in the two figures bellow. Both figures have four plots of the validation quantifiers (on the left) versus the number of principal components (nLV) used in the model. Plot a) shows the standard deviation ratio (SDR); plot b) shows the root mean squared error (RMSE); plot c) shows the correlation coefficient (R coeff) and plot d) shows the percentage of samples predicted with an absolute error smaller than 1 ºBrix (%error (º<1)).

In Figure 8 the validation quantifiers for the calibration (internal validation) are presented and Figure 9 presents the quantifiers for cross-validation.

It can be noted in Figure 8b that the RMSE decreases as the number of LVs increases, this is expectable since the model is attuning to the samples used for calibration. As for Figure 9b, the RMSE decreases till the 8th LV. After the 8th LV the error starts to increase, which means that the model is overfitted to the calibration data. This is in accordance to the precautionary warnings about overfitting stated previously.

By the results presented in Figure 9b, eight LVs could be used for prediction since it is the number that achieves the smallest error. However, it is important to remember that the purpose of the model is to make predictions with independent data. So, choosing a smaller number of LVs might result in a more robust model.

Analysing Figure 9d it can be observed that for the 8th LV, despite having a smaller RMSE, the result in percentage of error <1º is worse than for the 7th or 4th LV. This can be interpreted as, four LVs can be enough to make a good prediction. Another interpretation is

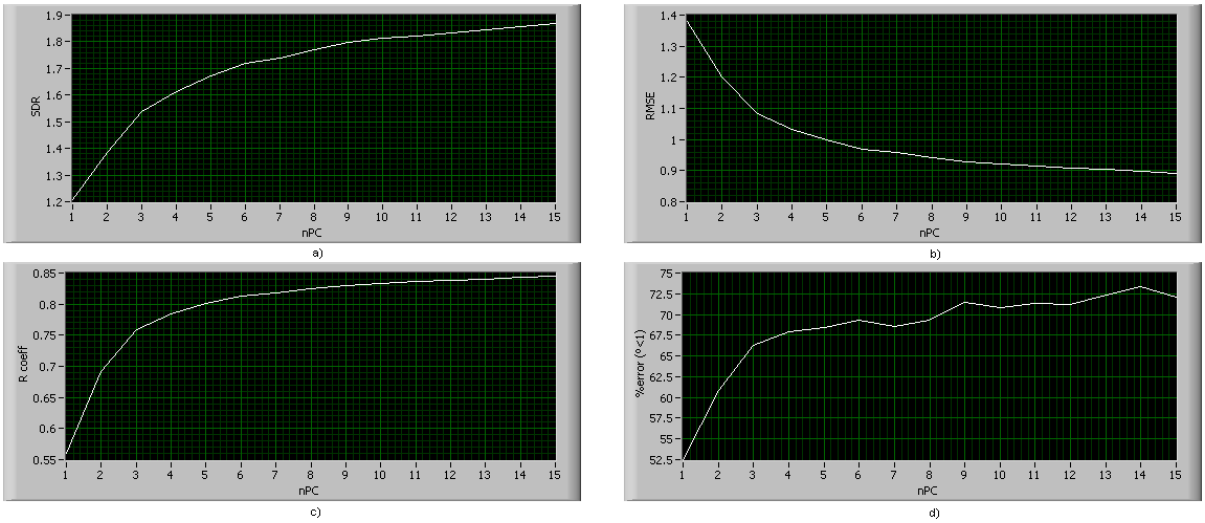


Figure 8 – Example of internal validation quantifiers

possible. The fourth graph quantifies a coarse measure: a prediction is accepted if it fails by less than 1 ºBrix. Since the measure is coarse, there is more room to incorporate components before overfitting is attained. Indeed, the best value for %error (º<1) is attained for 11 PCs, a maximum that is delayed by three PCs relatively to the other graphs. In terms of practical application, if an error of 1 ºBrix is acceptable, then one may consider increase the number latent variables relatively to the number obtained by observing the RMSE graph.

From the validation quantifiers presented in Figures 8 and Figure 9 different interpretations and choices of number of LVs to use can be made. So, it is uncertain what is the best choice for the number of LVs to use. It is important to continuously validate the model with external samples to make the correct choice.

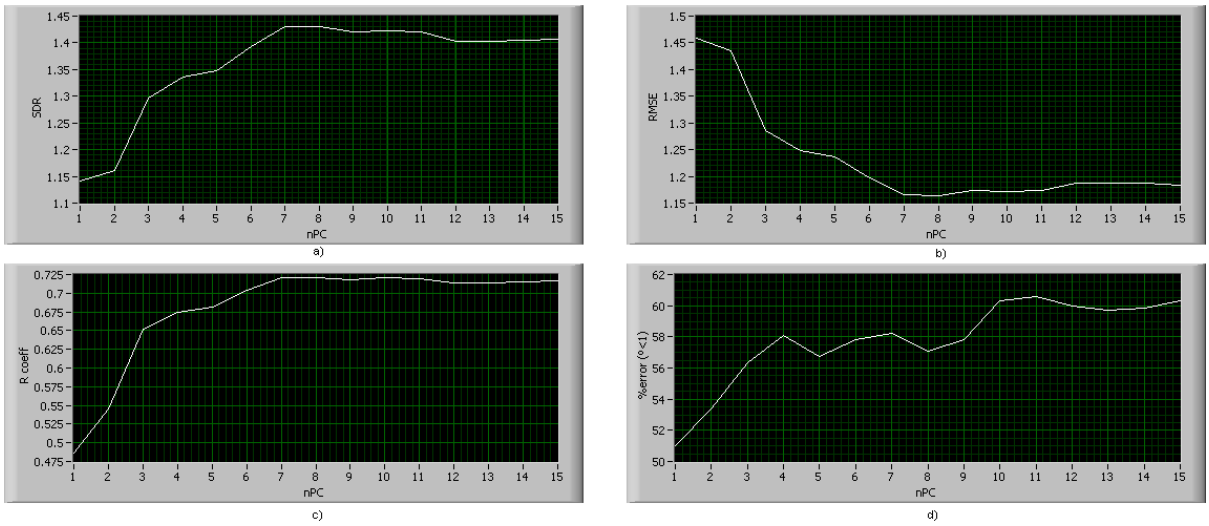


Figure 9 – Example of cross-validation quantifiers

4. Hardware and Measurement Setup

The hardware used for the spectrum measurements consists of:

- spectrometer;
- trigger;
- halogen light source and
- optical fibre cables.
- lenses

Additionally a computer is used for communication with the spectrometer, saving data, processing and presenting results. A thermometer was also used to monitor the setup temperature stability.

Spectrometers typically have the internal configuration presented in Figure 10.

The entrance slit is an aperture that allows light to get into the spectrometer. Normally an optical fibre cable is used to guide light into this entrance. A collimating mirror is used to collimate the light that spreads from the entrance slit and guide it to the transmission grating. In the transmission grating light spectrum is dispersed into different wavelengths. The grating diffracts each wavelength at a specific angle and a bundle of parallel rays are created for each wavelength. The focusing mirror focuses each bundle to a specific area of the image sensor because rays travelling at the same angle are focused in the same point of the focal plane. The sensor array is located in the focal plane capturing the wavelengths from the shortest to the longest. After that the accumulation of charge is converted to an electric signal. In our

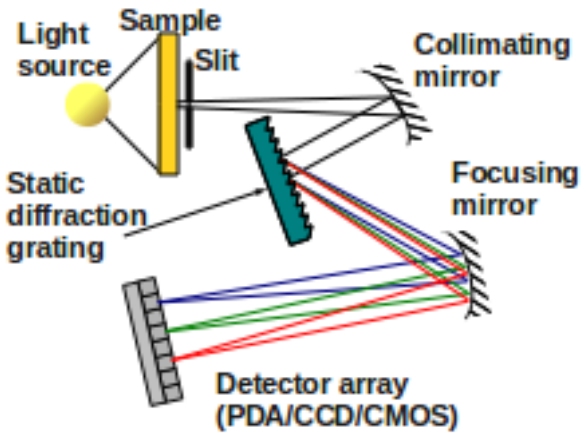


Figure 10 – Spectrometer internal configuration [44]

spectrometer the sensor used for conversion of the light signal into an electrical signal is a charge-coupled device (CCD). Specifically it is a linear image sensor (*Hamamatsu S10420-1006*).

The spectrometer is an *Hamamatsu C9405CA* from the family *TG-SWNIR CCD*. This spectrometer has a spectral response range from 500 to 1100nm with a spectral resolution of 5nm and an analogue to digital converter of 16 bits. The integration time may be changed from 10ms to 10000ms, which is the time the CCD has to accumulate charge.

Hamamatsu supplies proprietary drivers for communication between the computer and the spectrometer, and additional software is also provided for development of applications.

A trigger was created to sense the presence of a fruit in the calibration line. This allows the acquisition of spectrum to occur only when the fruit is in the region of interest. The trigger is especially useful when the line is moving. It is composed of a light source (UV-LED), a sensor (phototransistor) and proper circuitry for amplification and filtering. The circuit diagram is shown in Figure 11.

The light source is an *Ocean Optics HL-2000-FHSA-LL Tungsten Halogen Light*. It is suitable for application with wavelengths between 300nm and 2000nm. The light source features a long life light bulb (10,000 hours) and a fan for cooling. These features are important for stabilization purposes.

To guide the light from the light source to the fruit and from the fruit to the spectrometer, optical fibre cables are used. Specifically, a bifurcated fibre cable (*Ocean Optics QBIF600-VIS-NIR*) and a regular fibre cable (*Ocean Optics P1000-2-VIS-NIR*) are used. Both are suitable to be used with wavelengths from 400nm to 2100nm

The connection from the spectrometer to the computer is made with an USB cable. The

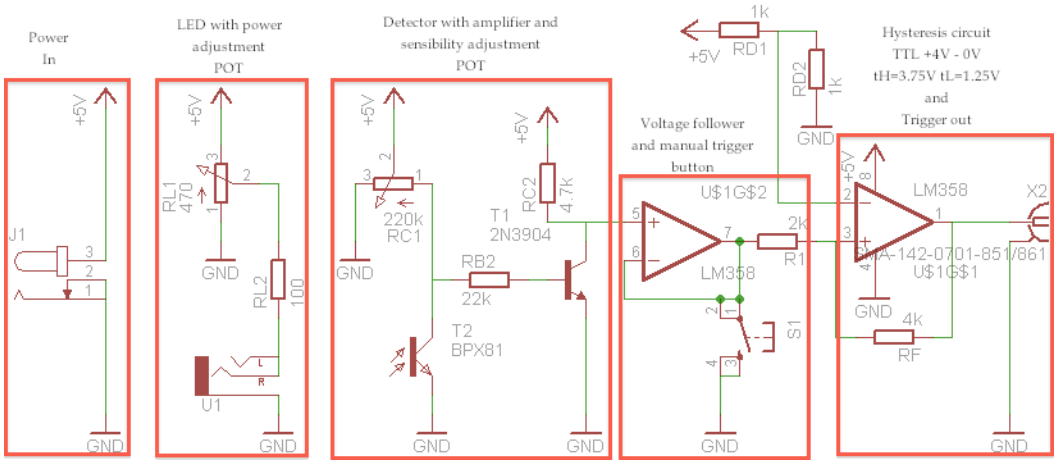


Figure 11 – Trigger circuit diagram

computer has an *AMD Sempron 3400+* (1.81GHz) processor and 896MB of ram.

For the temperature measurement a *Fluke Food Pro Plus* was used. It is based on an infrared sensor and has a resolution of ±1 °C between 0 °C and 65 °C. The thermometer was used to measure the temperature of the light source, the spectrometer, the environment and the *Teflon*® reference.

Figure 12 represents the measurement setup, which configures a geometry adapted to reflectance spectroscopy. A bifurcated optical fibre cable guides broadband light into the fruit. Light is focused on the fruit surface in two spots by the two focusing lenses attached to the fibres. Part of the light is reflected, and the other part penetrates the fruit, where it is absorbed and/or scattered. A small amount of backscattered light exits the fruit and is collected by the lens attached to the central optical fibre that guides it to the spectrometer. The spectrometer converts the light into an electrical signal and after to a digital signal that is sent to the computer. This setup was chosen after appropriate testing with different configurations made in a previous work.

Two steps precede the samples' spectrum acquisition. They are called dark and reference measurements. Dark measurement is performed with the light source turned off and with the environment light reduced to the minimum possible. The objective is to acquire the

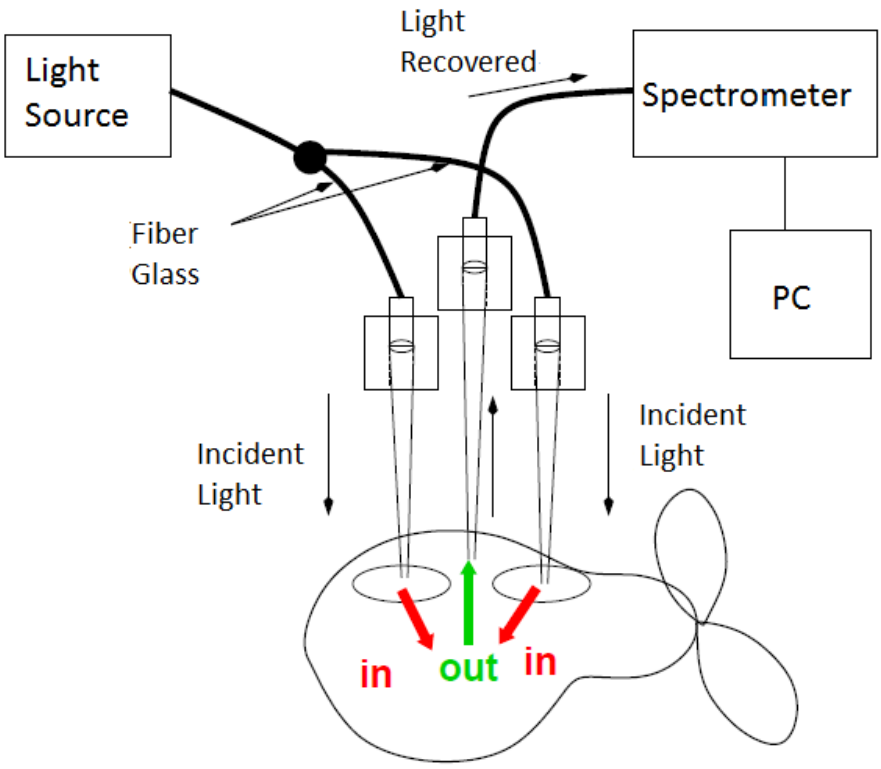


Figure 12 – Measurement setup for reflectance measurement

‘noise’. The dark spectrum is the result of environment light and instrumentation errors inherent to the spectrometer.

The basic mechanism of a CCD consists in the production of electrons from the incident photons. These electrons are stored in the pixel wells and then read out by the associated electronics. In an ideal CCD there would be no electrons accumulated in totally dark conditions. However, CCD's build up currents even in the absence of light because there are always electrons generated thermally. This is one of the major sources of noise. However, there are also other contributions, especially on-chip CCD read noise and off-chip CCD camera noise. The former is caused by the on-chip electronics, and results in an error on the number of electrons read from each pixel. The latter is generated in the circuitry subsequent to the charge to voltage conversion, including amplification stages and analogue to digital conversion. The dark spectrum is a good measure of all these effects and it also takes into account any response non-uniformity in the sensor and any source of parasitic illumination.

Reference measurement is made with the light source turned on. The objective is to acquire the spectrum of the backscattered light without any absorption. This serves to make a comparison with the samples' spectrum, since samples greatly modify the light spectrum. A *Teflon*® disc is used for the reference measurement. *Teflon*® has the desired properties, it scatters most of the light and absorbs a minimum. After these steps the samples' spectrum acquisition can be performed. The dark and reference steps are always necessary when the setup was turned on recently. If possible, dark and reference measurement should be performed regularly during measurements. This captures the variations on the environment and removes them from the samples' spectrum. Equation 13 presents the calculation to obtain a reflectance spectrum. The sample spectrum is subtracted by the dark and a ratio is performed between the result and the subtraction of the reference by the dark.

$$Reflectance = \frac{Sample - Dark}{Reference - Dark}$$

Equation 13

The measurements are influenced by the surrounding elements, like fluctuations in environment light or temperature. The temperature of the spectrometer and wear can also cause drift. To verify the stability of the spectrometer, measurements were performed to check the spectrometer temperature dependence. The spectrometer was set to acquire fifty times in a row, and this procedure was repeated with a rate of one measurement per minute. Along with the spectrum measurements, temperatures from the floor, the Teflon® reference, the spectrometer and the light source were taken once per minute. Laboratory lights were

turned off to avoid light variations. However, to simulate a real application scenario where there will always be some light entering the measurement chamber, the laboratory door was kept open. The pc screen was also turned on keeping the light in the laboratory to a fairly minimum. So, the conditions are similar to those described for dark measurements.

A plot of the temperature of the spectrometer and average counts of the dark measurements can be seen in Figure 13.

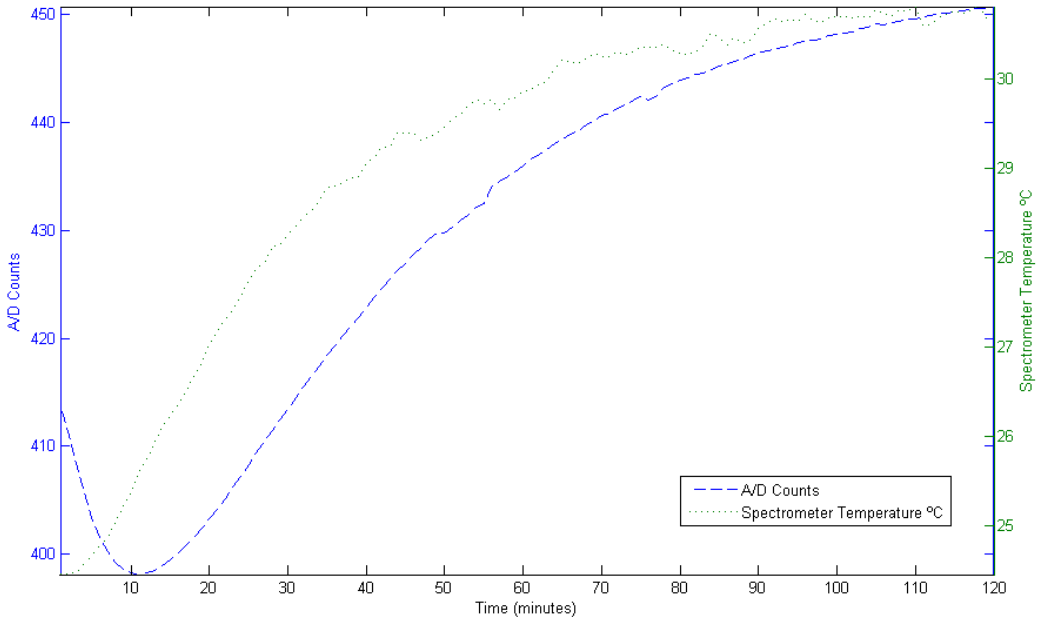


Figure 13 – Spectrometer temperature and average A/D counts

The right axis represents the spectrometer temperature in ºC, the left represents A/D counts averaged for all pixels and for the fifty measurements and the bottom axis represents time in minutes. In the figure it can be seen that after a decrease of the A/D counts (while the temperature was increasing), the counts start to rise. Temperature stabilizes near 31ºC and the A/D counts stabilize around 450, after 120 minutes. It is important to refer that this test started a little after the spectrometer was connected to the mains power. This is important because the starting temperature of the spectrometer varies if it is connected to the mains (hot start) or not (cold start). If the spectrometer has a cold start its temperature is close to the ambient temperature, while if it has a hot start its temperature is around 26ºC for a ambient temperature of 21ºC. Essentially it was noticed that the spectrometer will take around 120 minutes to stabilize if it has a cold start, compared to 20 minutes for a hot start. Several measurements were made in the conditions stated above, for different starting points (hot or cold start) and different room temperatures. This can be seen in Figure 14, were a plot of the temperature versus the average A/D counts is presented. The figure shows that there is a

relationship between the spectrometer temperature and the A/D counts (Dark Noise), with an optimum temperature at 23ºC.

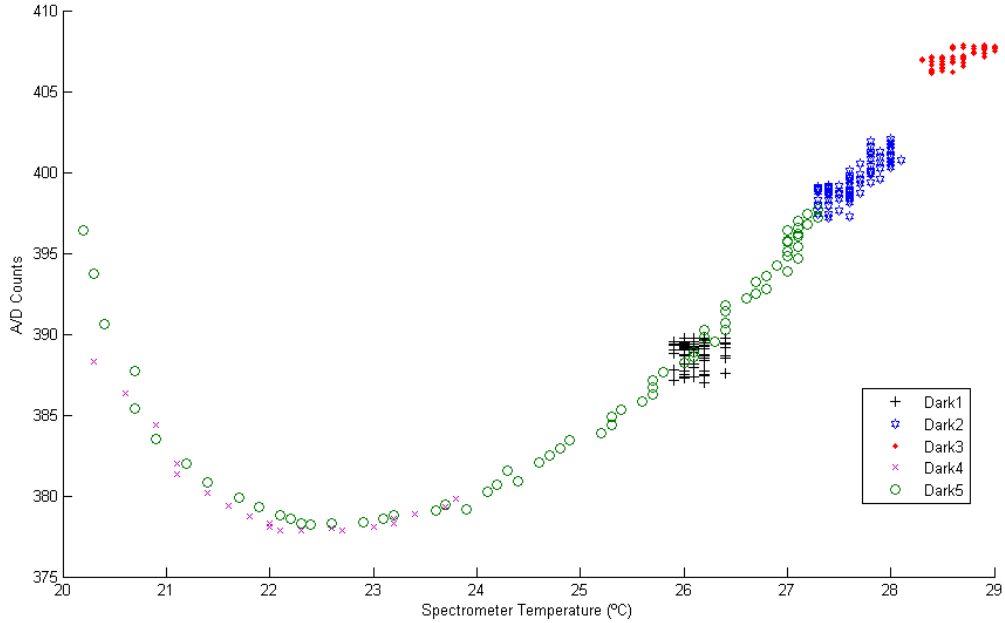


Figure 14 – Spectrometer temperature measurements over average A/D Counts

The light source was also tested for stability. After the spectrometer stabilization was achieved, reference measurements were taken. The light source was turned on at the measurements start. The measurements were performed in the same way as those for the spectrometer. Results showed that the light source stabilizes after 20 minutes. This can be seen in Figure 15. The figure shows that the A/D counts stabilize despite the temperature rise. The temperature also stabilizes after 30 minutes.

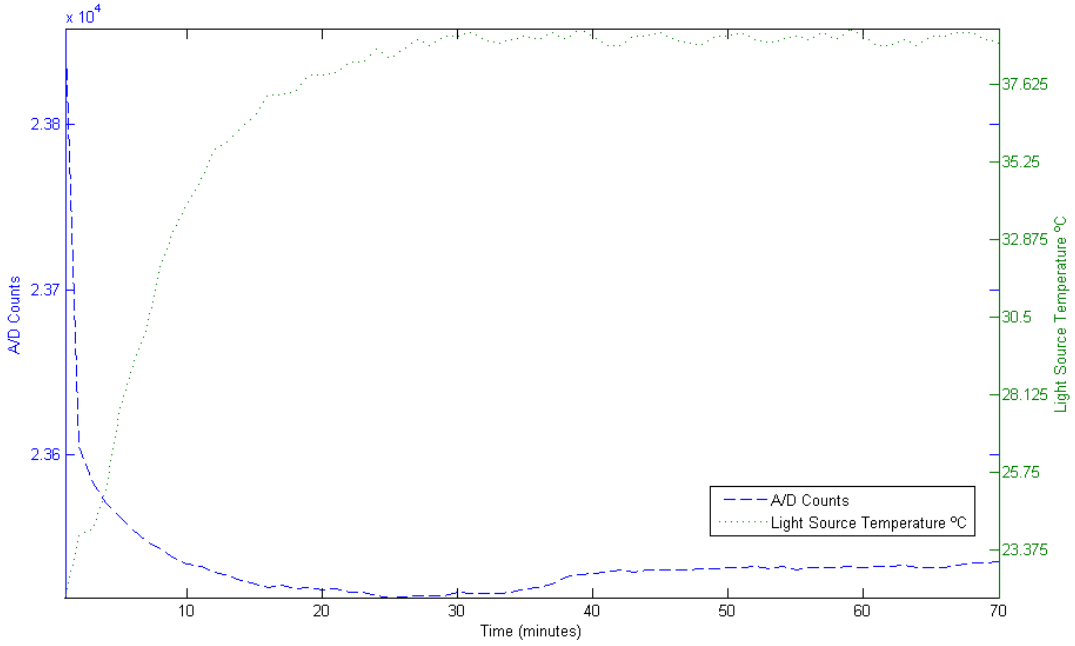


Figure 15 – Temperature of the *Teflon*® reference and average A/D Counts

5. Software Description

The aim of this work was to create a useful laboratory tool, to facilitate the spectra acquisition, model creation and model validation. For this purpose, the software was developed in LabVIEW™. LabVIEW™ has the advantage of generating automatically a graphical interface while creating the algorithms, with the needed variables and functions. In this chapter the user interfaces are presented and the various functionalities of each are described.

The software opening panel is the 'Main (Menu)'. Figure 16 shows the 'Main (Menu)', which is a launching panel. In the panel there are options to start the 'Data Acquisition (Mode 1)', to 'Create Models (Mode 2)' and to perform the 'Model Validation (Mode 3)'. There is also an exit button to close the software.

The three modes of the software are described in the next sections.

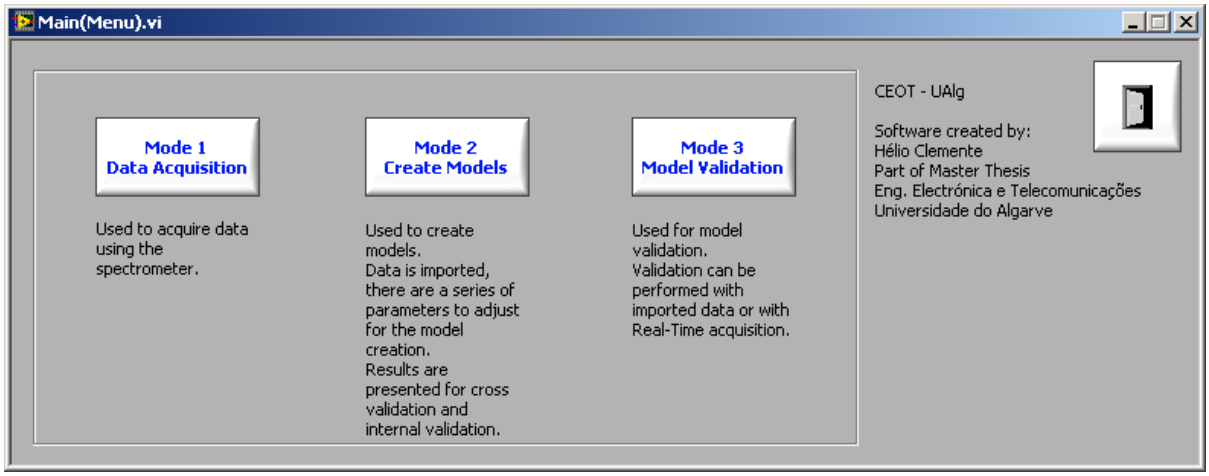


Figure 16 – Software Main (Menu)

5.1. Data Acquisition

The data acquisition mode or mode 1 was developed to control the spectrometer and to acquire, visualize and save the spectrum data. Normal operation in this mode include the functions:

- adjust the acquisition parameters;
- acquire data (with or without repetitions);
- save the acquired data to a file.

The 'Data Acquisition' panel is presented in Figure 17. The current acquisition parameters (integration time, trigger edge and trigger mode) are presented in the panel. There

is also a button (‘Adjust Acquisition Parameters’) to open the sub-panel ‘Spectrometer Control’ where the acquisition parameters can be set and tested.

The ‘Spectrometer Control’ sub-panel is presented in Figure 18. The panel has three indicators that are used to provide information to the user about the recognition of the acquisition device by the software, the creation of a data link connection to the device and whether the acquisition parameters were changed. The acquisition parameters that can be set are integration time, trigger edge and trigger mode.

As described before, the spectrometer uses a CCD sensor that converts photons to an electric signal. The conversion is performed during a period of time where the CCD accumulates a photoelectric signal, which is the integration time. The output signal level depends of the integration time. If the integration time is higher the signal level will increase.

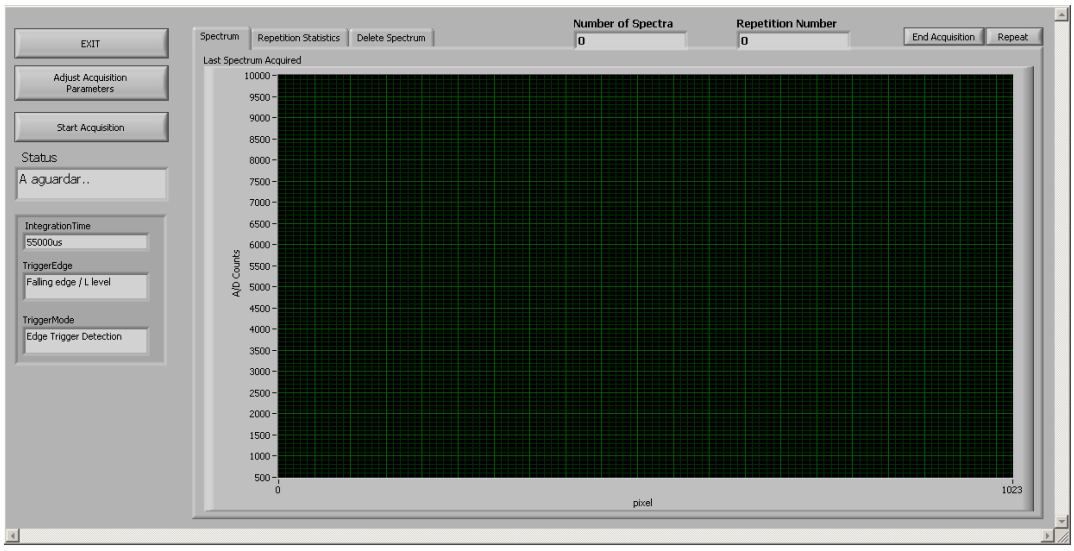
The trigger mode parameter has three options, internal trigger, edge trigger and gate trigger. This parameter affects what is the data string to be read by the computer. To understand that, let us recall once again that the spectrometer performs in sequence the accumulation of charge (during integration time period), its conversion to a digital signal and finally the charge clearing in each pixel. After this cycle the process starts from the beginning. This process is performed independently of the trigger mode the trigger mode, however, determines what and how the data signal is to be obtained.

If the trigger mode is set to internal trigger (free-run operation mode) the data obtained by the computer is the last available. In this mode a spectrum can be obtained for every integration time period.

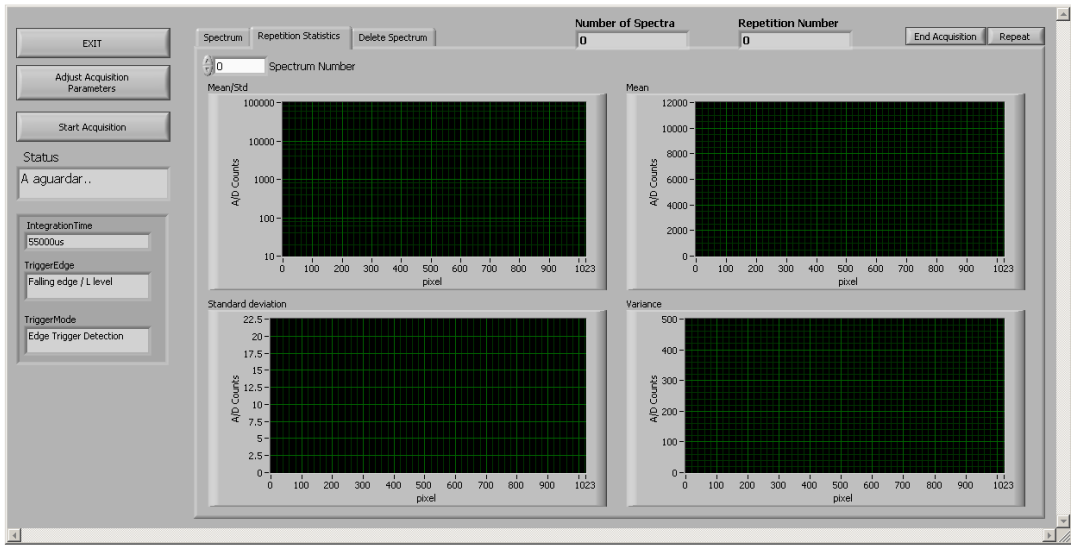
In trigger edge mode, the spectrometer holds the digital data after the trigger is on. This data is kept by the spectrometer until the computer reads the data or until a new trigger event. If the trigger is set on while data is being held by the spectrometer, then the held data is updated to the new data.

Trigger gate mode is similar to trigger edge mode. The difference is that while in trigger edge mode only one spectrum is held by the spectrometer, in trigger gate mode more than one spectrum may be held. The number of held spectra is the determined by the trigger set on and set off instants. While the trigger is on the data being accumulated is labelled, when the computer accesses the data from the spectrometer, only the labelled data is read.

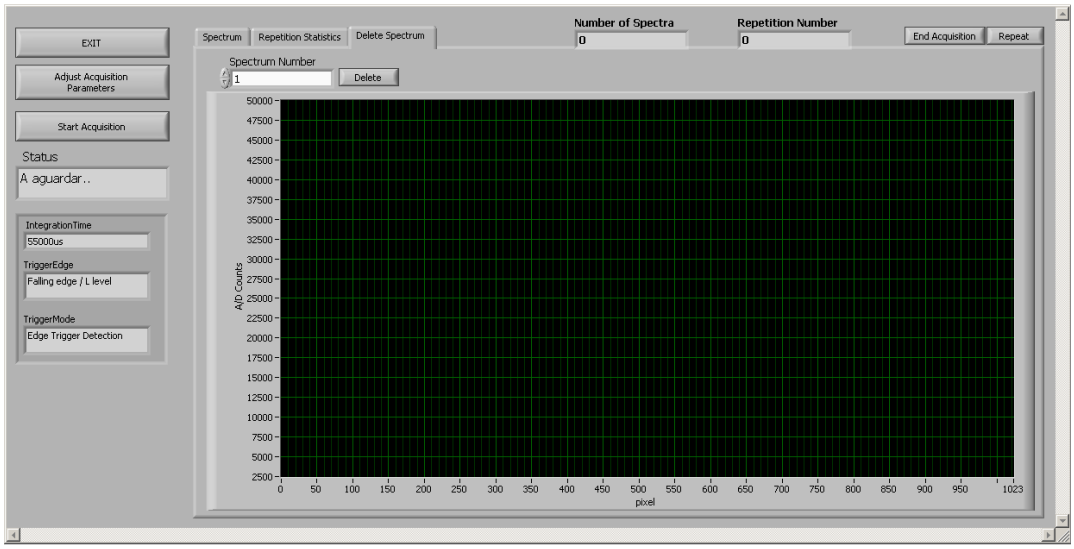
The trigger edge parameter has two options: ‘Rising Edge’ and ‘Falling Edge’, corresponding respectively to the choice of a rising or falling edge in the trigger signal as the trigger set on. The trigger edge parameter is used by the trigger gate mode and trigger edge mode.



a)



b)



c)

Figure 17 – Data Acquisition panel, a) Spectrum tab, b) Repetition Statistics tab and c) Delete Spectrum tab

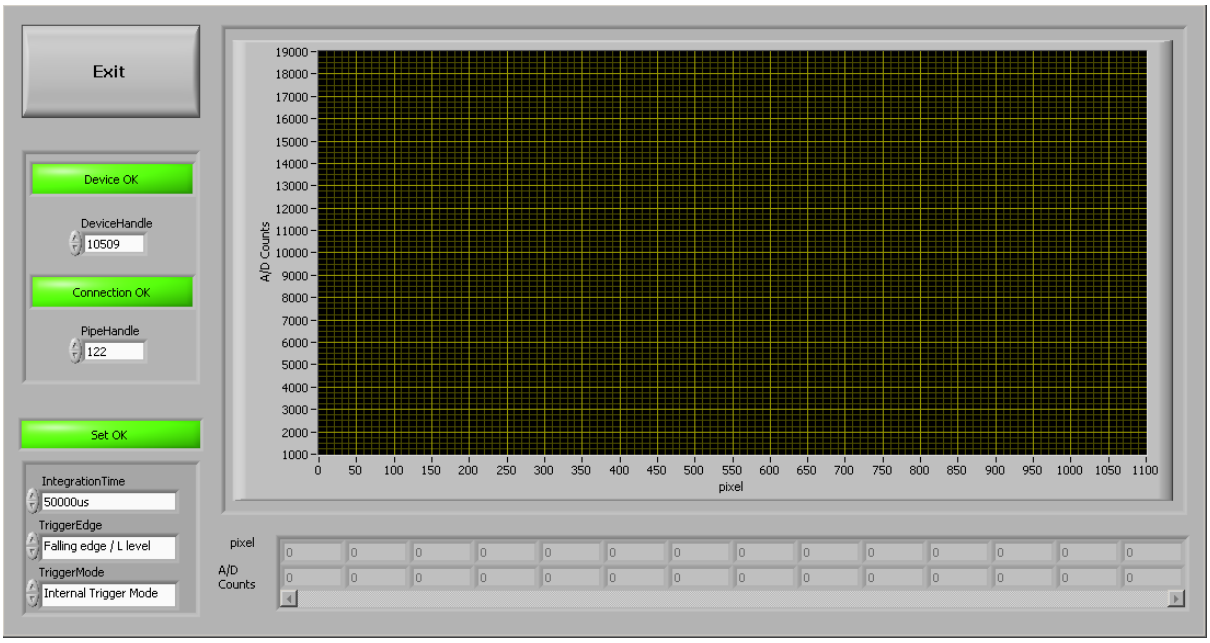


Figure 18 – Spectrometer Control sub-panel

In our setup the trigger is produced by a photogate constituted by a led and a phototransistor. The blockage of the light path by the fruit produces a series of dark/illuminated sequences at the phototransistor, translated into a sequence of 0 and 1 logic levels. The trigger edge could then be chosen between the two transition possibilities of light/dark or dark/light.

The software verifies periodically if there is new data in the spectrometer. The 'Spectrometer Control' sub-panel has a plot and a table to present the last acquired data to the user. This feature is useful to verify that the acquisition parameters meet the desired requirements for the subsequent measurements (for example, to check if the integration time is optimized, yielding a high but not saturated signal). If the settings are correct the sub-panel can be closed and the measurements can be started in the 'Data Acquisition' panel.

The user can start the acquisition pressing the button 'Start Acquisition' in the 'Data Acquisition' panel. When acquiring data, the last acquired spectrum is presented in the plot of the 'Spectrum' tab. The 'Spectrum' tab is presented in Figure 17a.

There are two more tabs in the panel. The 'Repetition Statistics' and 'Delete Spectrum' tab.

The 'Delete Spectrum' tab is presented in Figure 17c, this tab is used to delete undesired data. Undesired data is acquired when the user accidentally sets the trigger or when the sample was in an unwanted position. In the 'Delete Spectrum' tab, the 'Spectrum Number' control is used to navigate through the acquired spectra. The correspondent spectrum data is presented in the plot.

Figure 17b presents the ‘Repetition Statistics‘ tab. This tab is used when acquisition is performed with repetitions. Repetitions are useful to verify how the movement of the fruits affect the measurement, or how the fruit positioning changes the spectrum. Repetitions can be performed using the ‘Repeat’ button. The Fruits are introduced in the calibration line and the corresponding spectra are acquired. After that, the same fruits are set in the same order to perform the acquisition repetition. After the ‘Repeat’ button is pressed, the fruits can be passed again. The repetitions may be performed an arbitrary number of times. The statistics of the repetitions are presented for each fruit. There are four plots to visualize the repetitions statistics: ‘Mean/ Std’, ‘Mean’, ‘Standard deviation’ and ‘Variance’. The 'Mean/Std' plot shows the mean spectrum over the spectra standard deviation and it is equivalent to the concept of signal to noise ratio. Its value will be high for spectral variables with high and reproducible responses (high mean and low standard deviation) and will be low for spectral variables with low and/or non-reproducible responses (low mean and/or high standard deviation).

To end an acquisition, the ‘End acquisition’ button must be pressed. After the button has been pressed the acquisition ends and a sub-panel is presented to save the acquired data. This sub-panel is called ‘Save Spectrum’ and is presented in Figure 19.

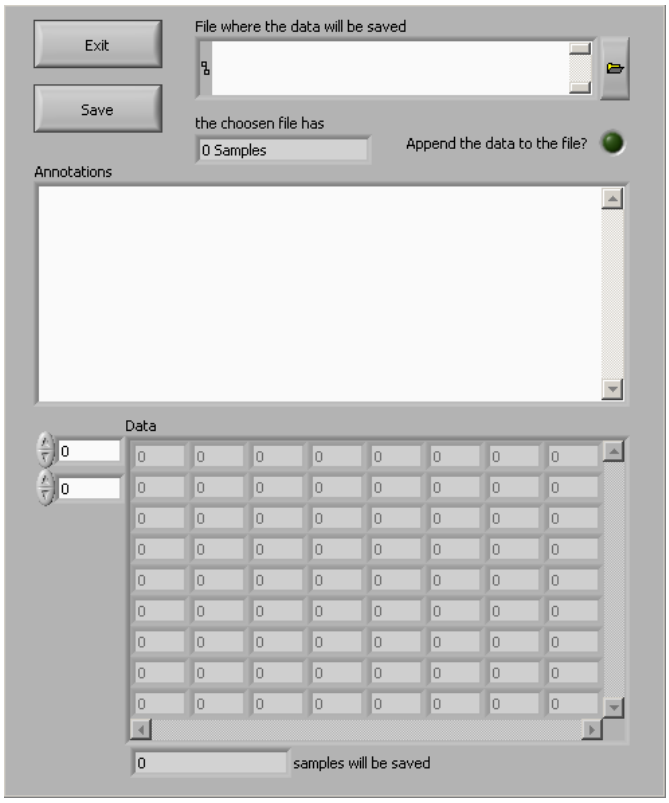


Figure 19 – Save Spectrum sub-panel

The ‘Save Spectrum’ sub-panel has a path control to choose the path and the name of the file to save the spectral data. The chosen file can be a new one or one of the already existing. In the latter case the option (‘Append the data to the file?’) may be activated to append the new data to the data already written in the file. If the option is not activated the newest data substitutes the data contained in the file. There is an ‘Annotations’ box to include observations about the data. These annotations are saved together with the spectrum data for further consultation. A ‘Data’ table is also shown. The acquired spectra data is presented on the ‘Data’ table and the user can use this table to review the data.

5.2. Create Models

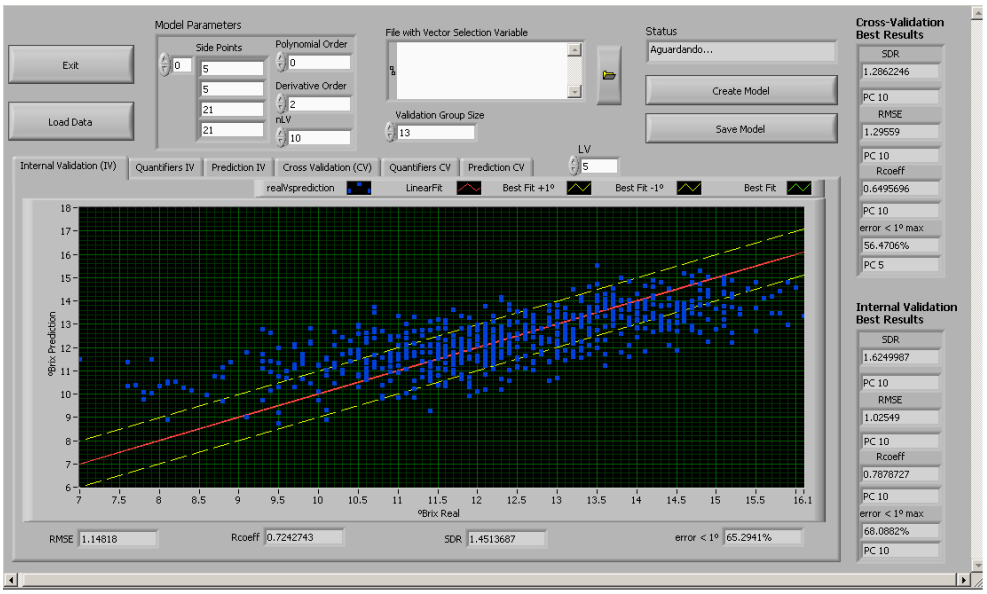
The create models mode or mode 2 was developed to create models and to validate the model using cross-validation and internal validation. In order to create a model it is necessary to:

- load data
- adjust the model parameters
- choose the size of the validation group

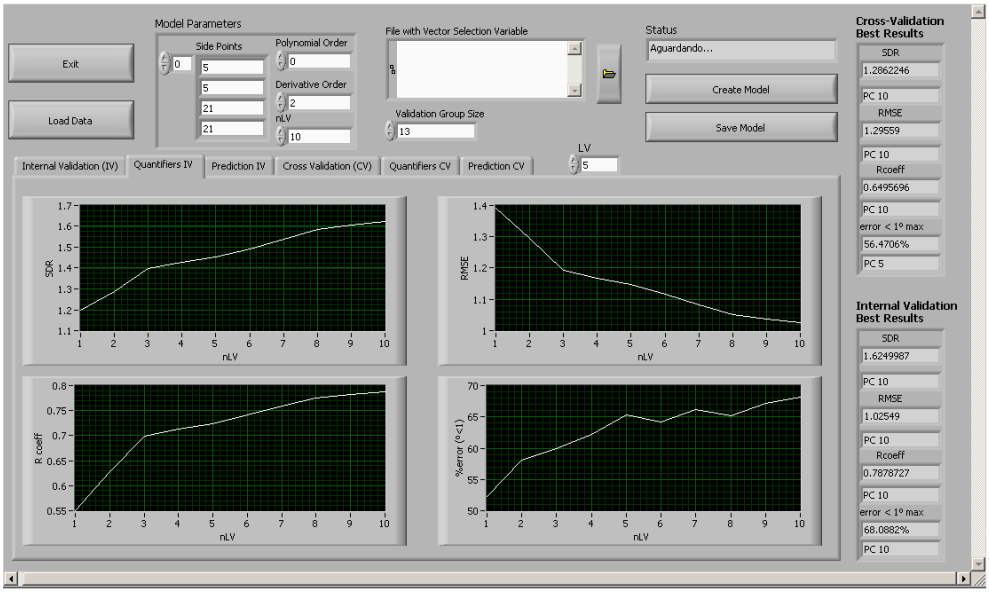
The ‘Create Models’ panel is presented in Figure 20. For better understanding the panel usage, in Figure 20 an example is presented with the parameters used for a model creation and the model validation results.

Within the ‘Create Models’ panel there is a section with the ‘Model Parameters’. The model parameters are defined by the spectrum preprocessing options and number of latent variables (‘nLV’) to calculate (recall from the section on Partial Least Squares that the latent variables correspond to the principal components retained in PLS). The methods available for spectral preprocessing include the Savitzky-Golay algorithm [30], the moving average smoothing, the derivation and the auto-scaling. Auto-scaling is always performed after the other methods. The preprocessing methods depend on six parameters, defined in the options ‘Side Points’ (the amplitude of the window used to calculate the average, expressed in number of points), ‘Polynomial Order’ (the order of the polynomial used to interpolate the data) and ‘Derivative Order’ (the order of the derivative applied to the data).

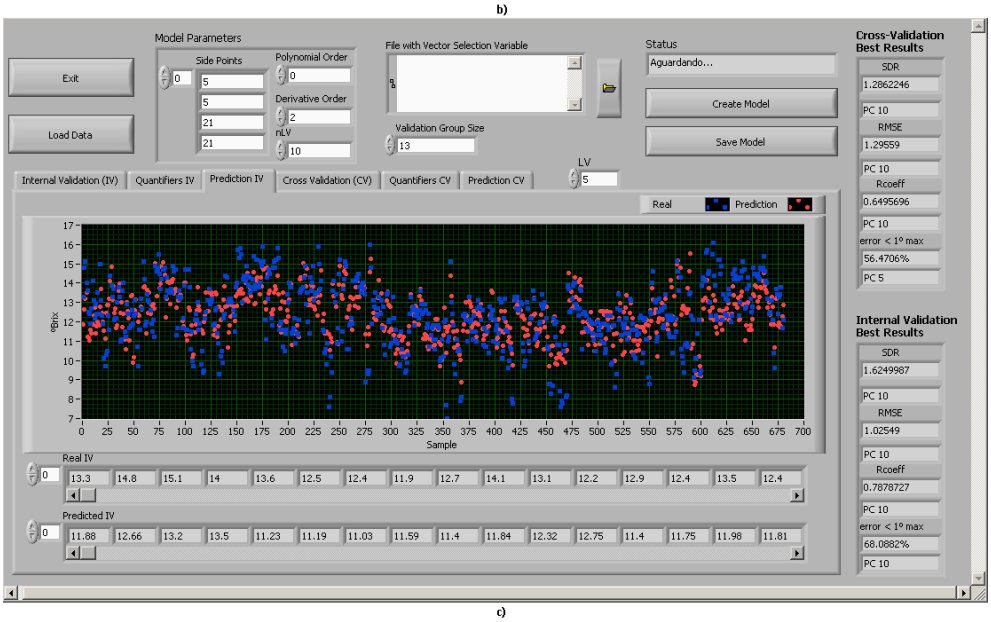
The Savitzky-Golay and the moving average are both smoothing algorithms. The latter corresponds simply to an average over a window of N neighbours around each point of the spectrum. The former is a preferred tool of spectroscopists and, for the same window, fits a polynomial of specified order around each point of the spectrum.


Model Parameters
Side Points
Polynomial Order
Derivative Order
nLV
Exit
Load Data
File with Vector Selection Variable
Validation Group Size
Status
Aguardando...
Create Model
Save Model
LV
Internal Validation (IV)
Quantifiers IV
Prediction IV
Cross Validation (CV)
Quantifiers CV
Prediction CV
realVsprediction
LinearFit
Best Fit +1°
Best Fit -1°
Best Fit
°Brix Prediction
°Brix Real
RMSE 1.14818
Rcoeff 0.7242743
SDR 1.4513687
error < 1° 65.2941%
Cross-Validation Best Results
SDR 1.2862246
PC 10
RMSE 1.29559
PC 10
Rcoeff 0.6495696
PC 10
error < 1° max 56.4706%
PC 5
Internal Validation Best Results
SDR 1.6249987
PC 10
RMSE 1.02549
PC 10
Rcoeff 0.7878727
PC 10
error < 1° max 68.0882%
PC 10

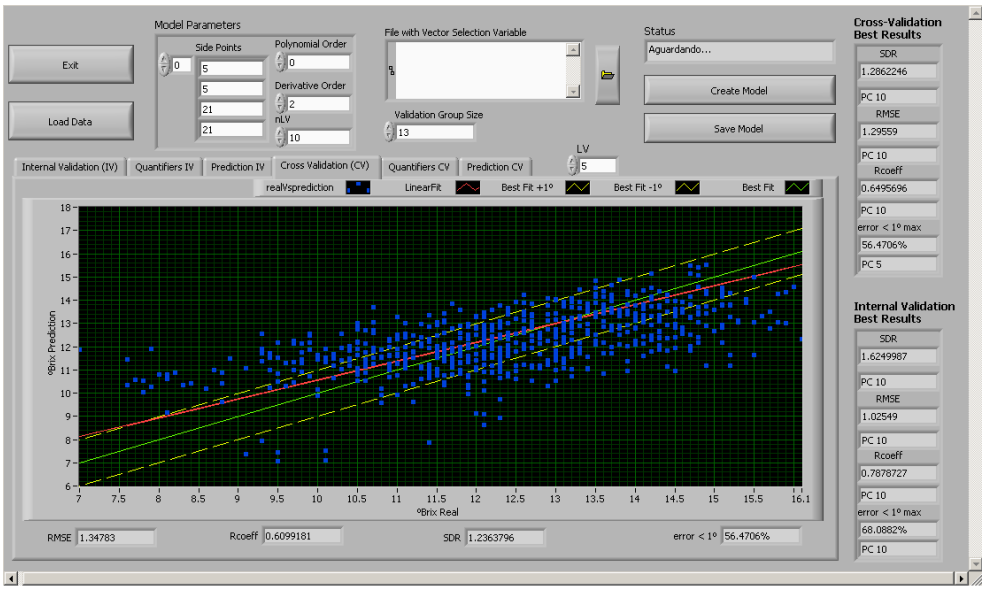
a)


Model Parameters
Exit
Load Data
Status
Aguardando...
Create Model
Save Model
Internal Validation (IV)
Quantifiers IV
Prediction IV
Cross Validation (CV)
Quantifiers CV
Prediction CV
SDR
RMSE
R coeff
%error (<1)
nLV
Cross-Validation Best Results
Internal Validation Best Results

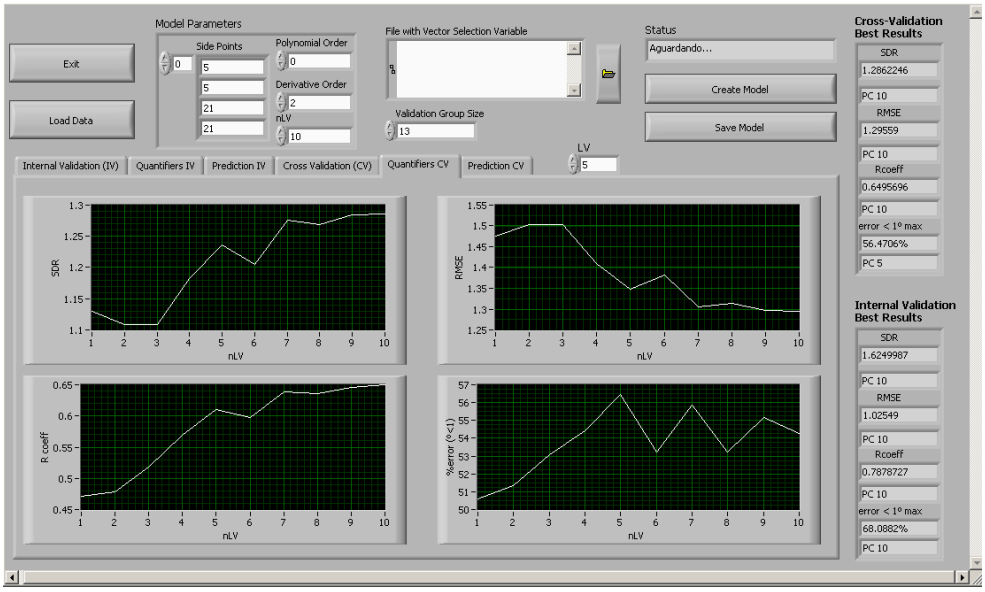
b)


Model Parameters
Exit
Load Data
Status
Aguardando...
Create Model
Save Model
Internal Validation (IV)
Quantifiers IV
Prediction IV
Cross Validation (CV)
Quantifiers CV
Prediction CV
Real
Prediction
°Brix
Sample
Real IV
Predicted IV
Cross-Validation Best Results
Internal Validation Best Results

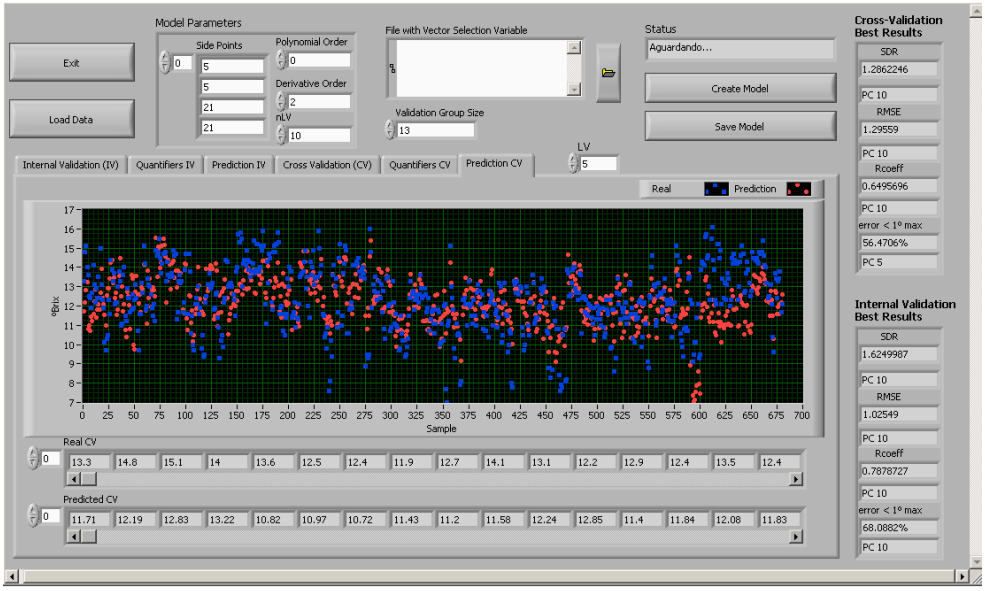
c)



d)



e)



f)

Figure 20 – Create Models panel, a) Internal Validation (IV) tab, b) Quantifiers IV tab, c) Prediction IV tab, d) Cross Validation (CV) tab, e) Quantifiers CV tab and f) Prediction CV tab

The Savitzky-Golay algorithm allows to perform smoothing and derivation at the same time, and so its use requires the specification of the three options 'Side Points', 'Polynomial Order' and 'Derivative Order'. To use the Savitzky-Golay algorithm the polynomial order must be greater then 0. If the polynomial order is set to 0, the moving average smoothing is used instead.

The size of the averaging window defined by 'Side Points' is calculated by Equation 14.

$$Window\ Size\ =\ (Side\ Points\ *2)\ +1$$

Equation 14

If the derivative order is equal to 0, no derivation is performed. When derivations are performed in combination with the moving average smoothing, the moving average is done twice before and after each derivation. The size of the moving average window can be different for each derivation.

The extra control 'File with Vector Selection Variable' is used to load a file containing a vector. This vector contains information about the spectral variables that should be excluded from the model due to excessive noise.

The 'Validation Group Size' is used to set the number of samples that are used in cross-validation.

The 'Create Models' panel has six tabs where the validation results are presented. Three of the tabs present results for the internal validation (IV) and the other three present results for the cross-validation (CV). Both IV and CV tabs have the same features and appearance and therefore only the cross-validation tabs are described next.

Cross Validation (CV)' tab (Figure 20d) shows a plot of the results for ºBrix prediction versus the real ºBrix for the chosen number of latent variables ('LV'). In this plot the data is represented by squared dots. The extra lines represent the data linear fit (red line); the perfect prediction ('Best Fit' green line, it is a straight line of slope 1, that is, predicted ºBrix= real ºBrix); and the perfect prediction shifted by an error of ± 1º (yellow dashed lines). These extra lines helps to visualize how the model behaves, specifically how close is the prediction from being perfect and whether the predictions fit within an acceptable error band of 1 ºBrix around the true values. Obviously, the prediction error is larger for the dots further away from the green line. On the 'Cross Validation' tab there are also indicators showing the results for the validation quantifiers ('RMSE', 'Rcoeff', 'SDR' and 'error < 1º' - see section 2.2.6). The results shown correspond to the number of LVs displayed in the 'LV' control. Changing the number of LVs updates automatically the model, the predictions and all the quantifiers.

Figure 20e shows the ‘Quantifiers CV’ tab. The tab has four plots with the validation quantifiers versus the number of latent variables. The description of these type of plots has been already made in section 2.2.6, when describing PLS model quantifiers and the Figure 8 andFigure 9.

The ‘Prediction CV’ tab (Figure 20f) presents a plot of the two arrays with the predicted and the real ºBrix values. Since the data is presented by sample number, the user may check for collective behaviour of sample groups; for example, detecting groups with particularly bad or good description by the model similarly to the ‘Cross Validation (CV)’ tab, the results shown corresponds to the number of LVs displayed in the 'LV' control.

The program calculates the quantifiers for each number of LVs, from 1 to ‘nLV’, for both internal and cross validation, and then identifies the best in each category. The indicators on the right side of the ‘Create Models’ panel (Figure 20) show these best quantifiers together with the corresponding number of LVs.

The first step to create a model using the ‘Create Models’ panel is loading the data. By pressing the button ‘Load Data’ (in Figure 20) the sub-panel ‘Load Data’ (Figure 21) is opened.

The data files have three components. One is the ‘central’ file (extension ‘.spec’) where the annotations are saved. Other file contains the spectral data (extension ‘.vind’) and the other contains the destructive data (extension ‘.vdep’). The three files have the same name and a different extension.

Figure 21 – Load Data sub-panel

In the sub-panel ‘Load Data’, the files are chosen using the path control ‘Path to file’. When a spectrum file is chosen, the annotations associated with the data are shown in the ‘Annotations’ box. The data file is ‘pre-loaded’ and a histogram of the destructive data (ºBrix) is shown in the tab ‘Statistics - File’.

The data loading is confirmed using the button ‘Load Data’. The ‘Load Data’ panel has one indicator showing the name of the loaded files and one with the number of loaded samples (both on the upper right of the panel). A histogram of the loaded destructive data is shown in the ‘Statistics – Loaded Data’ tab, similarly to the ‘Statistics - File’ tab.

The user can load as many files as needed. To add a new pre-loaded data file to the already loaded data, the control ‘Append Data From File’ must be switched on (as it is shown in Figure 21). Switching the control off replaces the (old) loaded data by the (new, pre-loaded) data.

Pressing the ’Exit’ button the panel is closed and the loaded data becomes effective in the calling panel.

It is possible to load an incomplete spectrum file (‘Path to file’). An incomplete file contains the spectral data but not the destructive data (measured ºBrix). The number of destructive data samples may also be in disagreement with the number of spectra samples. If this is the case, the sub-panel ‘Save Brix’ (Figure 22) is presented.

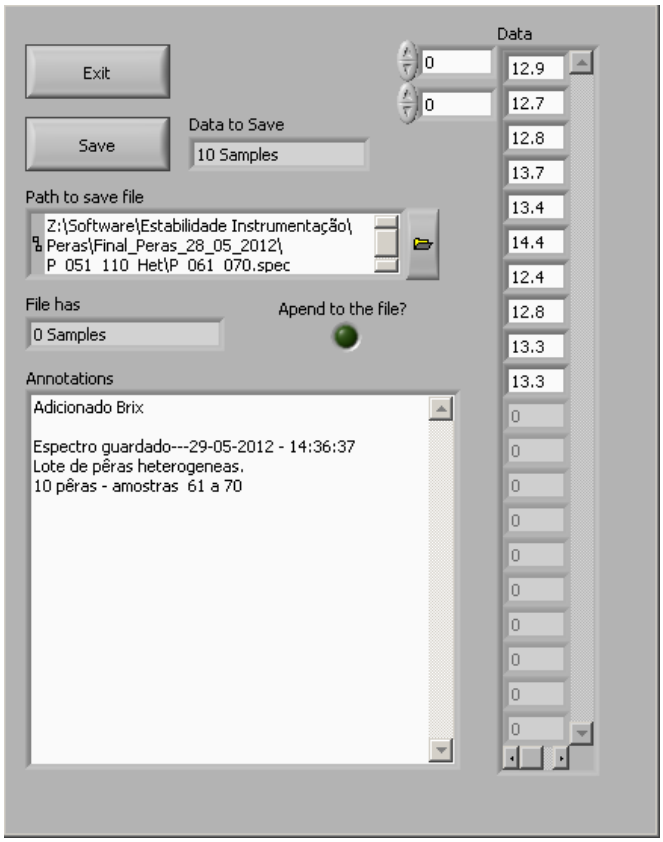


Figure 22 – Save Brix sub-panel

The ‘Save Brix’ sub-panel is used to input the destructive data. The ‘Data’ array presented is used for this purpose. The ‘Annotations’ box can be used to add more comments to the data. Using the ‘Append to the file’ control, the destructive data can be added to the existing one or replace it. The ‘Save’ button is used to save the destructive data and/or the annotations.

After loading the data, setting the model parameters and the validation group size, a model is created by pressing the ‘Create Model’ button (in figure 20).

The model can be saved using the ‘Save Model’ button, which opens the sub-panel ‘Save Model’ (Figure 23).

In the sub-panel ‘Save Model’, the ‘Path to file’ control is used to choose the target file for saving. Comments can be added to the model using the ‘Annotations’ box. The model parameters used to create the model are presented in the panel.

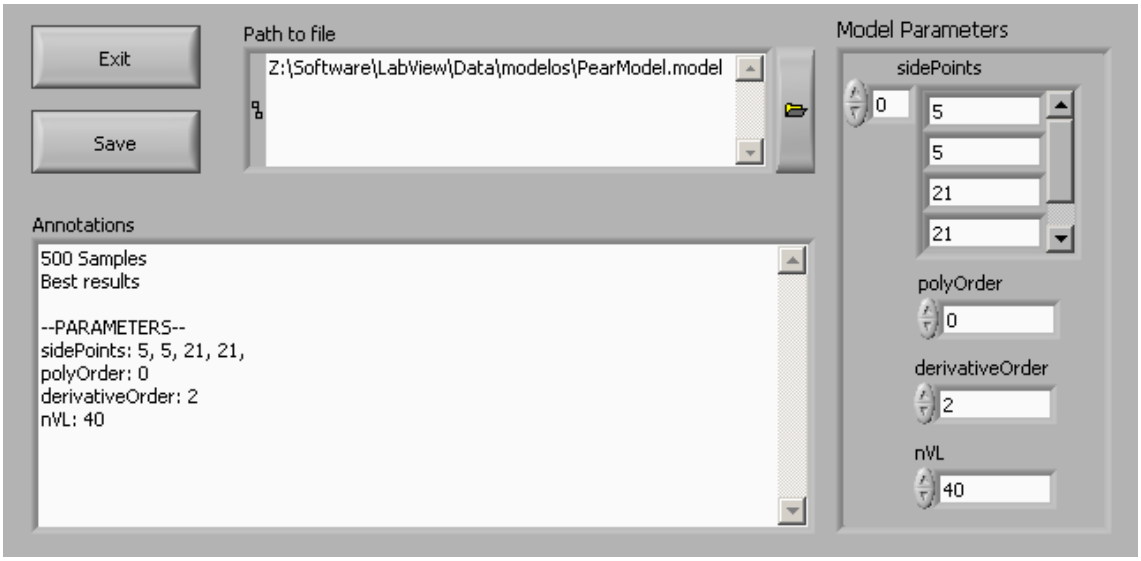


Figure 23 – Save Model – sub-panel

5.3. Model Validation

The model validation mode or mode 3 was developed, as the name implies, to validate the performance of the models created in mode 2. As previously stated, it is essential to verify the behaviour of the models when independent data is used. This is called external validation.

The ‘Model Validation’ panel (Figure 24) is aimed to execute external validations on previously built models. The external validation can be performed in real-time or using archived data. To do this, some functionalities of mode 1 and mode 2 are used. In fact, most of the functionalities of this panel were explained previously (in data acquisition mode and create models mode).

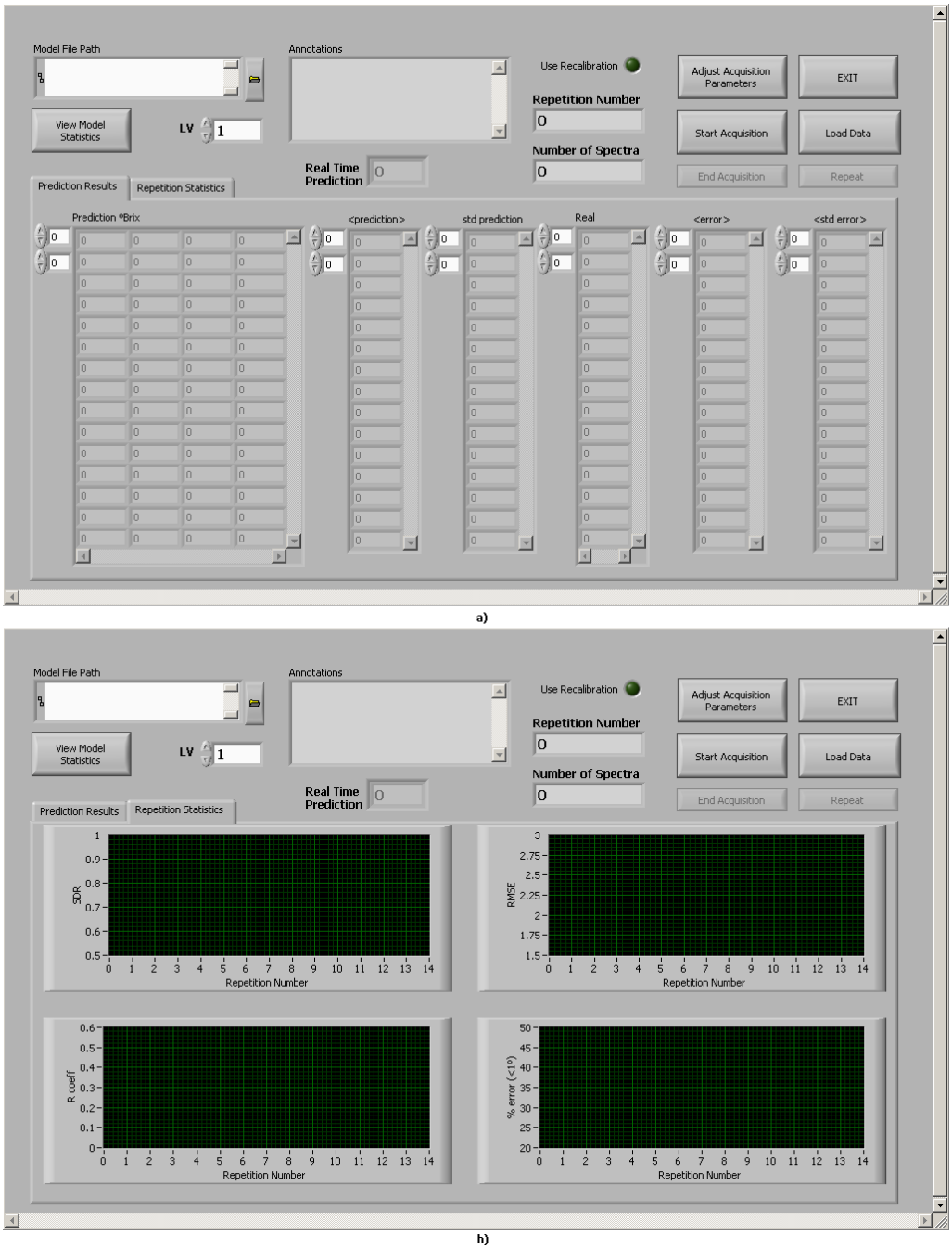


Figure 24 – Model Validation panel, a) Prediction Results tab and b) Repetition Statistics tab

The buttons 'Adjust Acquisition Parameters', 'Start Acquisition', 'End Acquisition' and 'Repeat' have exactly the same name and behaviour as those presented for the data acquisition mode.

The button 'Load Data' has the same name and behaviour of the 'Load Data' button in the create models mode (opens the 'Load Data' panel shown in Figure 21).

Despite the similarities to the previously presented modes, the model validation mode is used differently (because it has a different purpose). The succession of operations performed in this mode is:

- load a model
- chose the number of latent variables to use (only for the real-time prediction)
- acquire data (real-time prediction) or load data
- verify the results
- repeat if desired or load another data

The model is loaded using the control 'Model File Path'. When a model is loaded, the annotations of the model are presented in the 'Annotations' box. An image of the model statistics[5] (cross-validation quantifiers) can be viewed pressing the 'View Model Statistics' button.

When real-time prediction is used for validation, the number of latent variables to use is defined using the 'LV' control. The 'Start Acquisition' button is used to start real-time prediction. When a spectrum is acquired, a prediction is done for that sample and presented in the 'Real Time Prediction' indicator.

The results for the predictions are displayed in the 'Prediction Results' tab (Figure 24a). The table 'Prediction ºBrix' shows the results of the predictions, the lines corresponding to the samples and the columns to the repetitions. The average ('<prediction>') and the standard deviation ('std prediction') of the predictions (for each sample) are also displayed in the tab.

When the acquisition is ended, the sub-panel 'Save Spectrum' is presented to save the acquired data. After closing the sub-panel 'Save Spectrum', the sub-panel 'Save Brix' is presented. The 'Save Brix' sub-panel is used to input and save the destructive data. This data is passed back to the 'Model Validation' panel and presented in the 'Real' table. The error of the predictions is calculated and the average error ('<error>') and the error standard deviation ('std error') (for each sample) are presented in the 'Prediction Results' tab.

The 'Repetition Statistics' tab (Figure 24b) shows the validation quantifiers for the prediction results. This is similar to the validation quantifiers in Figure20b and 20e.

There are four plots with the validation quantifiers (STD, RMSE, R coeff and %error [<1º]) versus the repetition number. The plots show the results obtained for the quantifiers in each repetition round. Robust models will produce flat plots and less robust models will produce irregular curves. The number of LV may be changed in the control 'LV'.

When archived data is used to perform the validation, the results are displayed differently from those of real-time validation. When the data is loaded (using the sub-panel

[5] When a model is saved using mode 2, an image of the 'Create Models' tab (with the 'Quantifiers CV' tab selected) is also saved. So, the appearance of the model statistics image will be similar to figure 19e.

‘Load Data’), the prediction is performed for all the latent variables of the model. So, the prediction results are presented for each sample (lines) and for each latent variable (columns). This is the main difference between real-time and loaded data modes. In real-time mode the results take into account the repetitions and in loaded data mode the results take into account the latent variables.

6. Software Validation and Results

This chapter is intended to illustrate the utility of the developed software. The software was used to investigate two relevant questions appearing when one transposes a model build in the laboratory to the real environment of a calibration line:

- the need for model recalibration;
- the effect of the position of the fruit in the predictions.

The need for model recalibration stems from the fact that the model predictions may fail dramatically when applied to samples with characteristics very different from those of the fruits used to build the model. For example, if a model is built using only pears grown in a normal year, it will most probably fail in a year of drought. This is because the microstructure of the fruit tissue and its chemical composition changes, leading also to a change in the relation between the spectra and the ºBrix. However, it is found generally that the models retain the ability to distinguish between low and high ºBrix, although with meaningless absolute values (for example, negative values may be obtained). This means that it is possible to keep the models and look for mechanisms to correct the bias of the predictions.

Recall that the spectra and the destructive values are auto-scaled prior to the PLS algorithm. This means that the initial set of ºBrix values is transformed into a new set with zero mean and unit variance. Mathematically, the auto-scaled ºBrix values are

$$b' = \frac{b - \langle b \rangle}{\sigma}$$

Equation 15

where b represents the ºBrix value, $\langle b \rangle$ represents the average of the calibration ºBrix values, σ represents the standard deviation of the calibration ºBrix values and b' represents the auto-scaled ºBrix value. The predictions are made on the same scale. Therefore, an average ºBrix value is predicted as zero and a ºBrix higher (lower) than the average is predicted as a positive (negative) value. The final predictions are obtained by transforming back the auto-scaled predictions into the absolute scale by using the average and standard deviation of the population used in calibration:

$$b_{pred} = \langle b \rangle + \sigma b'_{pred}$$

Equation 16

where the subscript *pred* means predicted values.

The process of recalibration adopted in this work was the simplest possible. It assumes that the auto-scaled prediction is essentially correct (that is, b'_{pred} is correct) but that the

average and standard deviation of a given validation batch, $\langle b_{batch} \rangle, \sigma_{batch}$ may differ from those of the calibration set, $\langle b \rangle, \sigma$. The recalibration consists in estimating the average and standard deviation of the validation batch by destructive measurements of a few fruits (around 10) and make the prediction according to

$$b_{pred} = \langle b_{batch} \rangle + \sigma_{batch} b'_{pred}$$

Equation 17

The second problem investigated was the effect of the fruit orientation in the calibration line. Usually, a model is built on the laboratory by placing immobile and well-oriented fruits below the illumination/collection optics. In real processing, however, the fruits are randomly oriented relatively to the optics. A number of questions emerge: the predictions are affected by the position? How much? How is the model performance affected globally (in terms of its quantifiers)? On the other side, including random orientation in model construction could possibly improve its validation in real conditions. The software developed in this work was designed to easily handle repetitions and model creation through combination of different sample sources. Therefore it is appropriate to investigate the problem raised by the fruit random orientation.

The software was used in the three stages of the creation of a model: data acquisition, model creation and model validation. The procedures used and the results are described in the following sub-chapters.

6.1. Procedure

The software validation was performed using 'Rocha' pear. 'Rocha' pear is a Portuguese typical cultivar and it has unique characteristics. 'Rocha' pear is characterized by a typical russeting dispersed over the surface, an oblong shape, homogeneous pigmentation and a coloration that changes in shelf life from green to light yellow, depending on the maturity stage.

'Rocha' pear was chosen for three main reasons. First, it is one of the most important Portuguese export fruits, grown and commercialized by an important net of producers and cooperatives; secondly, it represents a challenge because of its characteristics of sudden change in colour and texture during shelf life; finally, because 'Rocha' pear was also used in The pears were chosen from different supermarkets in order to increase the fruit variability in terms of characteristics and origins. Therefore the pears could be used in different combinations to provide more robust models, and better validation results.

From each supermarket lot, two different sets of samples were created: an heterogeneous set (pears with different shapes, sizes, coloration and ripeness state and an homogeneous set (pears with uniform size and shape, very look alike).

In whole, there were four different sets and a total of 220 pears. Two heterogeneous sets with 50 and 60 samples each and two homogeneous with 50 and 60 samples each.

To ease the explanation of the spectra acquisition procedure and for clarification purposes, from now on the sets will be called 'Batch A' (50 heterogeneous pears), 'Batch B' (60 heterogeneous pears), 'Batch C' (60 homogeneous pears) and 'Batch D' (50 homogeneous pears).

The distribution of the ºBrix of the pears belonging to the four batches is depicted in Figure 25.

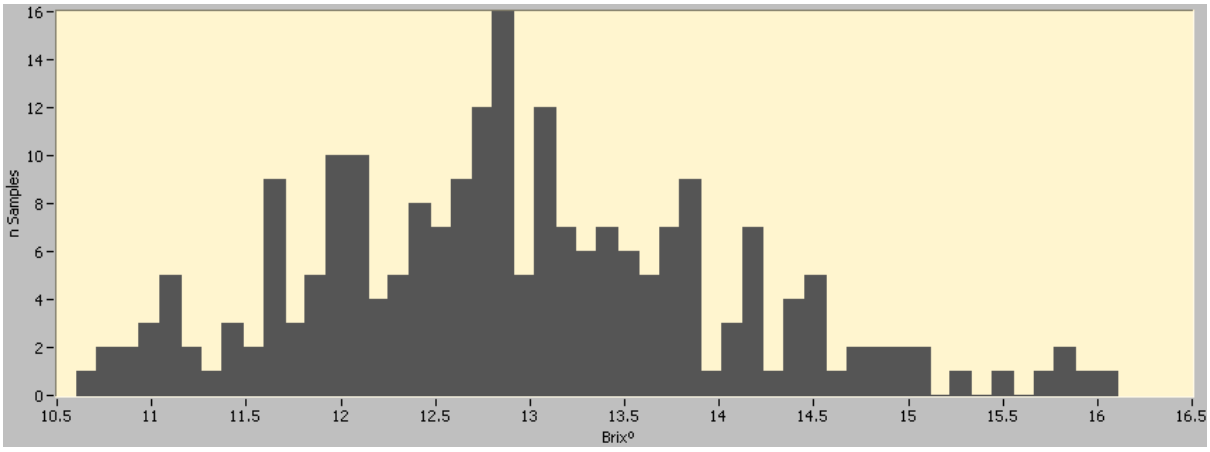


Figure 25 – Distribution of the ºBrix of the pears used in the tests

The spectra acquisition was performed using the software in the 'Data Acquisition' mode. All acquisitions were performed in the calibration line. The setup was positioned and adjusted in a way that took into account the possibility of future usage in the industry.

The spectrometer integration time used was equal for all the measurements. The choice of the integration time was made through a compromise between two parameters: the signal strength and the time effectively available to acquire the signal in a calibration line with moving fruits. On one side, the acquisition time has to be approximately half of the fruit transit time beneath the illuminated area (the areas around the stem and the calyx are not useful for prediction, representing approximately half of the transit time). On the other side, the signal strength should be as large as possible to increase signal to noise ratio (but not too close to saturation). Having both parameters into account the integration time was set to 55ms.

The Batch A spectra was measured with the calibration line stopped. One measurement was performed for each pear. It is important to remember that in reflectance spectroscopy,

dark and reference measurements must be performed. For Batch A the dark and reference measurements were performed before each pear measurement.

A typical pear spectrum is displayed in Figure 26.

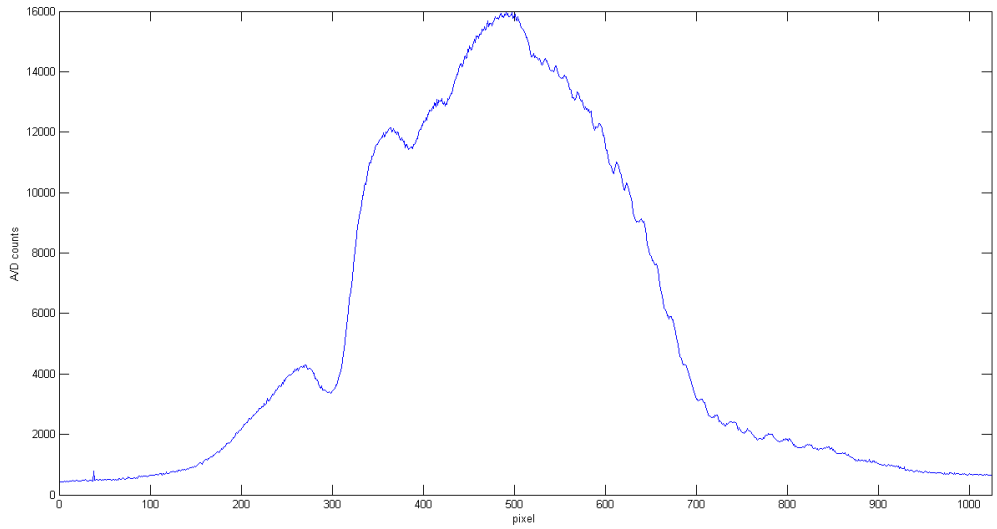


Figure 26 – Typical Pear Spectrum

For these measurements the trigger mode was set to 'internal trigger'. The measurements were performed sequentially, with the fruit aligned in the best position. Dark and reference measurements were also performed in this way.

Table 3 – Summary of the different data sets

Batch	N of pears	Uniformity	Motion	Group	N of pears	Repetitions	Position
A	50	heterogeneous	No	-	-	no repetition	best
B	60	heterogeneous	Yes	B1	10	10 for each pear, on different positions	random
				B2	50	no repetition	random
C	60	homogeneous	Yes	-	-	no repetition	best
D	50	homogeneous	Yes	-	-	10 for each pear, on different positions	random

The measurements for Batch B, Batch C and Batch D were made with the pears in motion (using the trigger to determine where the acquisition should start). The spectrometer parameters were the same for the three batches. The acquisition time was set to 55ms (the same time as Batch A) and the trigger mode was set to trigger edge.

The pears from Batch B were separated into two groups (the pears for each group were randomly chosen). The first group (B1) was constituted by 10 pears and the second group (B2) by 50 pears.

The first group of pears (Batch B1) was measured with repetitions, meaning that the spectrum for the 10 pears was measured sequentially (one pear after the other). The sequence was repeated 10 times (10 pears * 10 repetitions making a total of 100 spectra). The pears were placed in the calibration line in a random position (stem could be pointing to any side).

The second group of pears (Batch B2) was measured sequentially (50 spectra, one for each pear). The pears were positioned randomly.

The Batch C spectra were also measured sequentially (60 spectra, one for each pear), but these pears were laid flat, with the stem pointing backwards, which corresponds to the best orientation.

The Batch D was measured with repetitions (similarly to Batch B1), using all the pears (50 pears * 10 repetitions, totalling 500 spectra). The pears were positioned randomly.

After all the measurements were performed, the ºBrix for each pear was measured (destructively) using a digital refractometer.

The gathered spectra were used to verify the recalibration efficiency and how the pears position affects the predictions. The results will be presented in the next sub-chapter with the explanation on how the spectra were used to create the models and to perform external validations.

6.2. Results

The results presented below had the following objectives:

- verify the recalibration efficiency
  - using a homogeneous batch (Validation 1.1)
  - using a heterogeneous batch (Validation 1.2)
- verify how the pears position affect the predictions
  - observation of the prediction fluctuation (Validation 2.1)
  - test the use of repetition measurements fin calibration for the improvement of model robustness (Validation 2.2)

All the models created (for each of the validations) had three model parameters in common, the ‘Side Points’, the ‘Polynomial Order’ and the number of latent variables used.

The ‘Polynomial Order’ was set to 0. This means that a running average was used for smoothing the spectra.

The ‘Side Points’ was set to [5, 5, 21, 21]. This setting was used previously in the laboratory with good results.

The maximum number of latent variables (nLV) used to create each model was set at 30. This value is certainly larger than the optimal nLV and also large enough to identify the critical number of latent variables above which overfitting takes place.

The 'Derivative Order' was the only model parameter that was varied from 0 to 2.

Each of the validation results is presented below, in a separate table. The tables show the validation quantifiers corresponding to the best results (that is, correspond to the nLV that produce the best results). The best nLV is also shown together with the quantifier in a compact notation (e.g. 1.30 @LV2 meaning that the best result was 1.30 using 2 latent variables). The tables presented next display the results of validations for three different models.

Table 4 – Summary of the combinations used in the tests

Test	Calibration set	Simple Validation set	Validation with recalibration sets
Recalibration efficiency (validation with an homogeneous batch)	30 pears from batch C + A+B1 (one repetition only)+B2+D (one repetition only)	the remaining 30 pears from batch C (homogeneous)	from the 30, 10 used for recalibration and 20 for validation
Recalibration efficiency (validation with an heterogeneous batch)	A+B1 (one repetition only)+C+D (one repetition only)	B2 (heterogeneous)	from the 50, 20 used for recalibration and 30 for validation
Effect of the position of the pear on the predictions	A+B2+C+D (one repetition only)	4 pears from B1 (10 spectra per pear)	-
Randomized model to minimize the effect of pear positioning	D (all repetitions)	C	-
	D (one repetition only)		

6.2.1. Recalibration Efficiency

To verify the recalibration efficiency using a homogeneous batch (validation 1.1), 30 spectra of 'Batch C' were used (validation set 1.1). The remaining spectra from 'Batch C' and 'Batch A', 'Batch B1', 'Batch B2', 'Batch D' were used for the calibration (calibration set 1.1). In the cases of 'Batch B1' and 'Batch D' (where repetitions exist) only the spectra from the first repetition measurement were used. Three models were created using the calibration set 1.1.

All the combinations of batches used to obtain calibration and validation sets in all the four tests are presented in Table 4.

The validation set 1.1 was divided into 2 sets, 1 set used for recalibration (10 spectra) and the other used for validation (20 spectra). To verify the recalibration efficiency the validation was performed using recalibration/validation (10/20 spectra) and validation only (30 spectra).

The results obtained in this test are presented in Table 5. The results show that the use of recalibration does not improve the results. The best result of the standard deviation ratio (SDR) is 1.64 using 9 latent variables and derivative order 0 (no derivative).

The obvious drawback of recalibration is that it is based on a very small sampling to estimate averages and standard deviations. In this case only 10 samples were used. This under sampling may produce useful results if the validation and calibration sets are very different, and if, correspondingly, the predictions are very disparate from reality. In this case, under sampling may be enough to redirect the predictions towards an acceptable trend. However, if calibration and validation are not very different, the under sampling may introduce noticeable errors in the average and/or standard deviation. The result is the one displayed in Table 5: validation with recalibration results are worse than simple validation results.

The validation performed to verify the recalibration efficiency using a heterogeneous batch (validation 1.2) was very similar to validation 1.1 presented previously. The main difference is that data set used for validation was 'Batch B2' (validation set 1.2). In this case the entire batch was used (50 spectra) for validation. The remaining spectra from 'Batch A', 'Batch B1', 'Batch C' and 'Batch D' were used for the calibration (calibration set 1.2). In the cases of 'Batch B1' and 'Batch D' (where repetitions exist) only the spectra from the first repetition measurement were used. Three models were created using the calibration set 1.1.

The validation set 1.2 was divided into 2 sets, 1 set used for recalibration (20 spectra) and the other used for validation (30 spectra). To verify the recalibration efficiency the validation was performed using recalibration/validation (20/30 spectra) and validation only (50 spectra).

The validation results obtained in the heterogeneous case are presented in Table 6. The results are clearly different from those obtained in the homogeneous case. In the latter, simple validation was clearly the best approach while in the former both validation methods yield approximately equivalent results.

This is in agreement with the previous arguments: recalibration tends to be useful when calibration and validation sets are very different. In this second test the validation set is more

likely to present significant differences relatively to the calibration set, because it was chosen in such a way to be the more heterogeneous as possible.

Table 5 – Results for Validation 1.1 (Recalibration efficiency using a homogeneous batch)

Model Derivative Order	Validation Quantifier	Validation Using Recalibration	Validation Only
0	SDR	1.30 @LV2	**1.64 @LV9**
	RMSE	0.68 @LV2	0.60 @LV9
	R coeff	0.75 @LV10	0.79 @LV9
	% error (<1º)	85% @LV2	93.3% @LV7
1	SDR	1.03 @LV1	**1.50 @LV6**
	RMSE	0.86 @LV1	0.66 @LV6
	R coeff	0.46 @LV3	0.74 @LV6
	% error (<1º)	70% @LV1	90% @LV3
2	SDR	1.15 @LV1	**1.42 @LV20**
	RMSE	0.77 @LV1	0.69 @LV20
	R coeff	0.54 @LV11	0.72 @LV20
	% error (<1º)	75% @LV1	86.7% @LV19

Table 6 - Results for Validation 1.2 (Recalibration efficiency using heterogeneous batch)

Model Derivative Order	Validation Quantifier	Validation Using Recalibration	Validation Only
0	SDR	**1.28 @LV24**	**1.28 @LV24**
	RMSE	0.84 @LV24	0.79 @LV24
	R coeff	0.68 @LV13	0.65 @LV24
	% error (<1º)	76.7% @LV28	82% @LV24
1	SDR	**1.39 @LV6**	1.29 @LV12
	RMSE	0.78 @LV6	0.79 @LV12
	R coeff	0.74 @LV8	0.66 @LV7
	% error (<1º)	76.7% @LV8	86% @LV12
2	SDR	1.20 @LV7	**1.35 @LV20**
	RMSE	0.90 @LV7	0.75 @LV20
	R coeff	0.59 @LV7	0.67 @LV22
	% error (<1º)	76.7% @LV10	88% @LV20

This simulation suggests that recalibration may become clearly the preferred method in the presence of batches with very specific properties. In real application the user must decide if recalibration is worth or not. To do that, a sampling of around 10 pears in each new lot (for example, a lot from a new producer, a new lot coming from the storage cameras, etc.) should be always performed and a comparison between simple validation and validation with recalibration made. The decision is then taken based on this comparison.

6.2.2. Effect of the Position of the Pears in the Prediction

To verify how the position of the pears in the calibration line affects the predictions, two validations were performed. The first validation was performed to observe the effect of the random positioning of the pears. The second validation was performed to minimize the effect (prediction fluctuation) of the random positioning of the pears.

To observe the effect of the pears position, 4 pears and their corresponding 10 spectra were selected from 'Batch B1'. The 10 repetitions of 'Batch B1' were used, making a total of 40 predictions (10 for each pear). A model was created using 'Batch A', 'Batch B2', 'Batch C' and 'Batch D'. From 'Batch D' only the first repetition spectra were used. Thus, the calibration set was composed of 210 spectra. Only one model was created (derivative order 0) to simplify the analysis.

Figure 27 shows the prediction results for the 4 pears and for each of the 10 repetition spectra. The ºBrix predictions are displayed with a 'o' marker and the ºBrix real value is marked with a '*' marker. Different colours are used to distinguish the pears. It is noticeable that there is a large variation in the predictions for the same pear. Pear 'A' has a real Brix of 12.8º and presents the smallest variation with the predictions varying from 12.1º to 13.3º (amplitude of 0.6 ºBrix). Pear 'C' has a real Brix of 12.2º and presents the largest variation with the predictions varying from 11.1º to 13.6º (amplitude of a.3 ºBrix). So it is obvious that the pears positioning has an important impact in the prediction.

These results show that the positioning of the fruit in the calibration line may actually constitute the major source of error in the prediction of ºBrix. The best results presented above have shown prediction errors (RMSE) of the order of 0.7 ºBrix. On the other side, the errors derived from positioning are independent from the latter. Hence, one may estimate the total error e_{tot} as

$$e_{tot} = \sqrt{RMSE^2 + e_{pos}^2}$$

Equation 18

where e_{pos} is the error induced by positioning. Estimating $e_{pos} \approx 1$ ºBrix from our results, the total uncertainty in the predictions is raised to 1.2 ºBrix.

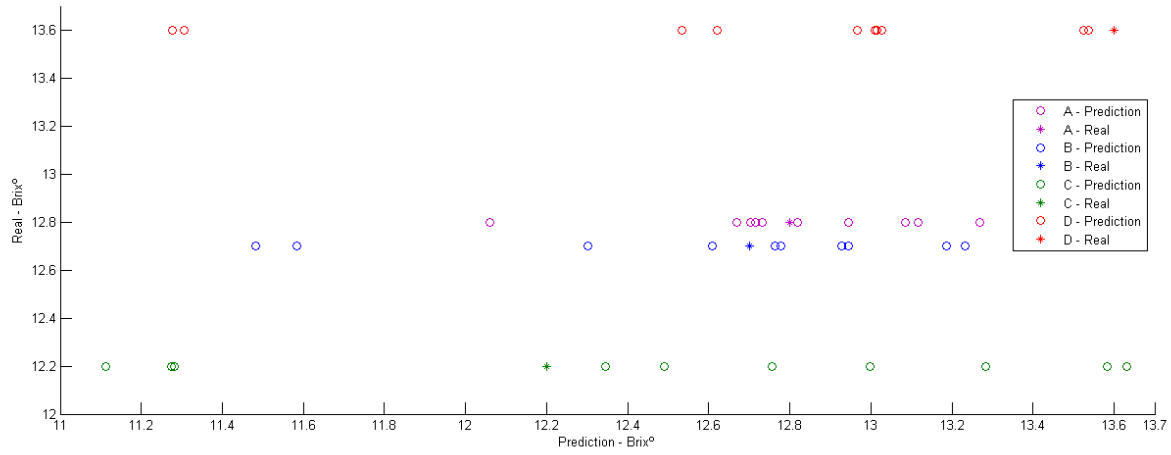


Figure 27 - ºBrix prediction for 4 pears and for each of the 10 repetitions spectra

A possible way to minimize the problem introduced by the fruit positioning is to understand how the shape of the spectrum is affected by the orientation of the fruit surface receiving the light. Then a transformation could be devised, in order to compensate for that effect. But this is clearly a new problem, outside the objectives of this work.

The final test aimed to verify the possibility of minimizing the effect of the random positioning of the pears by creating a "randomized" model. To create the models, the derivative order was varied (similarly to the previous validations results) and 2 different calibration sets were used (3 derivative orders * 2 calibration sets, in a total of 6 different models).

The calibration sets used were 'Batch D' using all repetition measurements (All Spectra) and 'Batch D' without repetitions (1st Repetition Spectra). The validation set used was 'Batch C'.

The results on Table 7 show that the models perform better when all the spectra with repetitions 'All Spectra' are used, contrasting with using only the first repetition '1st Repetition Spectra'. Although the models are only fairly 'good' (the best SDR was only 1.23), it is demonstrated that the use of repetition measurements can create a better model. In fact, the use of 'All Spectra' provides better results for the three derivative orders.

By including spectra for the same pear taken in different positions, the randomized model accommodates better the variations induced by positioning. It should be noted that batch C was measured in the best position. The difference in the two calibration sets is that

the randomized model gives a better answer for all positions, including the best, whereas the usual, non randomized model, has a worse response for all positions, including the best position. In brief, model randomization by acquisition of spectrum repetitions is a recommended procedure to improve model robustness.

Table 7 – Results for the minimization of the pears random position effect

Model Derivative Order	Validation Quantifier	All Spectra	1st Repetition Spectra
0	SDR	**1.23 @LV12**	0.74 @LV13
	RMSE	0.70 @LV12	1.16 @LV13
	R coeff	0.73 @LV12	0.75 @LV11
	% error (<1º)	91.7% @LV12	65% @LV13
1	SDR	**1.18 @LV3**	1.13 @LV5
	RMSE	0.73 @LV3	0.76 @LV5
	R coeff	0.71 @LV3	0.74 @LV16
	% error (<1º)	83.3% @LV3	81.7% @LV5
2	SDR	**1.00 @LV30**	0.72 @LV21
	RMSE	0.86 @LV30	1.18 @LV21
	R coeff	0.68 @LV30	0.72 @LV25
	% error (<1º)	80% @LV18	65% @LV20

7. Conclusions

ºBrix is one of the most important parameters for the evaluation of fruit internal quality. Several non-destructive methods for its determination have been tested in the laboratory with good results. However, the application of these methods on automatic calibration lines still poses major challenges.

Non-destructive methods are based on mathematical prediction models originated in a time-consuming calibration stage relating non-destructive data (the reflectance spectra in this work) and the corresponding destructive data (the ºBrix measured from refractometry in this work). The limitation of the laboratory procedures does not allow to simulate the real conditions of an automated fruit grading line.

The main goal of this research was to develop a system able to bridge the gap between laboratory and calibration line, by providing agile and versatile tools for spectra acquisition, creation of models and their validation. The system consists of a hardware component (spectrometer, optics, grading line prototype, pc) and a software component, written in LabVIEW™, to create and manage the models in a user-friendly interface. The creation of models is performed through the application of Partial Least Squares regression, one of the multivariate statistical techniques with more proven results. All the algorithms for data analysis were transposed to LabVIEW™. These specific features of the system allow fast tests and real-time evaluation of the prediction models results. Furthermore, they allow to anticipate specific problems in the calibration line and to search for possible solutions.

The system was tested by investigating two problems with practical relevance in industrial, real life applications. The first problem was to determine the utility of model recalibration when analysing batches of fruits with characteristics that may differ significantly from those of the fruits used for calibration. The second problem was the quantification of the error induced in the predictions by the random orientation of the fruit in the calibration line. In connection with this question, we have compared the advantages of creating models from fruits with randomized positions versus models created from fruits in aligned positions.

The first problem was tackled through the simplest method of recalibration, known as bias correction, whereby the mean and the standard deviation of the calibration population are substituted by the corresponding values of the current batch under analysis. This method was evaluated and the results show that the method is plausibly useful for heterogeneous

batches of fruit. However, more tests are needed to investigate further the advantages of recalibration.

Concerning the second problem, the prediction error introduced by the random positioning of the fruit relatively to the collection optics has been quantified and shown to be about 1 ºBrix. To reduce this source of 'noise' a different form of calibration was attempted through the inclusion in the calibration matrix of repetition spectra (from the same pears). The results show that this method can be used to improve the models' robustness in grading line real conditions.

Summarizing, in this work the one has implemented a system for non-destructive measurement of fruit internal quality through diffuse reflectance spectroscopy. The system was conceived and built from the basic principles. The system allowed to demonstrate the feasibility of the method and was also designed in such a way that it could simulate the real conditions of an automated grading line. At the same time it is based on a user-friendly interface, allowing the user to build his own models and to evaluate them through a battery of statistical tests. Our system may be updated in the future with more statistical techniques for model calibration and validation, making it an open tool for future developments.

8. Bibliography

[1] "Food and Agriculture Organization of the United Nations - for a world without hunger." [Online]. Available: http://faostat.fao.org/. [Accessed: 23-Mar-2012].

[2] F. M. da Silva and A. M. M. B. de Morais, *Boas Práticas de pós-colheita para Frutos Frescos*, 1ª edição. Orgal, 2000.

[3] D. Almeida, *Manuseamento de Produtos Hortofrutículas*, 1ª edição. Principia, Publicações Universitárias e Científicas, 2005.

[4] M. Knee, Ed., *Bases biológicas de la calidad de la fruta*, 1ª ed. Acribia S.A., 2008.

[5] A. Cavaco, R. Guerra, and D. Antunes, "Errar ou não errar: como é que os métodos não-invasivos vão mudar a avaliação da qualidade de frutos e legumes?," *Frutas legumes e flores*, pp. 46–47, 2012.

[6] P. Butz, C. Hofmann, and B. Tauscher, "Recent developments in noninvasive techniques for fresh fruit and vegetable internal quality analysis," *Journal of food science*, vol. 70, no. 9, pp. 131–141, 2005.

[7] P. Armstrong, "Nondestructive acoustic and compression measurements of watermelon for internal damage detection," *Applied Engineering in Agriculture*, vol. 13, no. 5, pp. 641–645, 1997.

[8] T. Sun, K. Huang, H. Xu, and Y. Ying, "Research advances in nondestructive determination of internal quality in watermelon/melon: A review," *Journal of Food Engineering*, vol. 100, no. 4, pp. 569–577, Oct. 2010.

[9] J. a. Abbott, "Quality measurement of fruits and vegetables," *Postharvest Biology and Technology*, vol. 15, no. 3, pp. 207–225, Mar. 1999.

[10] M. Taniwaki, T. Hanada, M. Tohro, and N. Sakurai, "Non-destructive determination of the optimum eating ripeness of pears and their texture measurements using acoustical vibration techniques," *Postharvest Biology and Technology*, vol. 51, no. 3, pp. 305–310, Mar. 2009.

[11] K. Kato, "Electrical Density Sorting and Estimation of Soluble Solids Content of Watermelon," *Journal of Agricultural Engineering Research*, vol. 67, no. 2, pp. 161–170, Jun. 1997.

[12] B. A. Snopok and I. V. Kruglenko, "Multisensor systems for chemical analysis: state-of-the-art in Electronic Nose technology and new trends in machine olfaction," *Thin Solid Films*, vol. 418, no. 1, pp. 21–41, Oct. 2002.

[13] F. K. Che Harun, J. A. Covington, and J. W. Gardner, "Portable e-Mucosa System: Mimicking the biological olfactory," *Procedia Chemistry*, vol. 1, no. 1, pp. 991–994, Sep. 2009.

[14] S.-M. Kim, P. Chen, M. J. McCarthy, and B. Zion, "Fruit Internal Quality Evaluation using On-line Nuclear Magnetic Resonance Sensors," *Journal of Agricultural Engineering Research*, vol. 74, no. 3, pp. 293–301, Nov. 1999.

[15] R. Guerra, I. V. Gardé, M. D. Antunes, J. M. da Silva, R. Antunes, and A. M. Cavaco, "A possibility for non-invasive diagnosis of superficial scald in 'Rocha' pear based on chlorophyll a fluorescence, colorimetry, and the relation between α-farnesene and conjugated trienols," *Scientia Horticulturae*, vol. 134, pp. 127–138, Feb. 2012.

[16] M. Ciscato, M. Sowinska, M. van de Ven, F. Heisel, T. Deckers, J. Bonany, and R. Valcke, “Fluorescence imaging as a diagnostic tool to detect physiological disorders during storage of apples,” in *Acta Horticulturae, 553*, 2001.

[17] J. Belasque, Jr., M. C. G. Gasparoto, and L. G. Marcassa, “Detection of mechanical and disease stresses in citrus plants by fluorescence spectroscopy,” *Applied Optics*, vol. 47, no. 11, p. 1922, Apr. 2008.

[18] V. Leemans, H. Magein, and M.-F. Destain, “On-line Fruit Grading according to their External Quality using Machine Vision,” *Biosystems Engineering*, vol. 83, no. 4, pp. 397–404, Dec. 2002.

[19] T. Brosnan and D.-W. Sun, “Improving quality inspection of food products by computer vision—a review,” *Journal of Food Engineering*, vol. 61, no. 1, pp. 3–16, Jan. 2004.

[20] B. Nicolai, K. Beullens, E. Bobelyn, a Peirs, W. Saeys, K. Theron, and J. Lammertyn, “Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review,” *Postharvest Biology and Technology*, vol. 46, no. 2, pp. 99–118, Nov. 2007.

[21] K. H. Norris, “Design and development of a new moisture meter,” *Agric. Eng.*, vol. 45, no. 370, 1964.

[22] A. M. Cavaco, M. D. C. Antunes, J. Marques da Silva, and R. Guerra, “A preliminary approach to the prediction of ‘Rocha’ pear skin pigments by a Vis/NIR reflectance spectroscopy,” *Acta Horticulturae*, vol. 858, pp. 373–378, 2010.

[23] A. Cavaco, P. Pinto, M. Antunes, J. Silva, and R. Guerra, “‘Rocha’ pear firmness predicted by a Vis/NIR segmented model,” *Postharvest Biology and Technology*, vol. 51, no. 3, pp. 311–319, 2009.

[24] A. M. Cavaco, D. Antunes, J. M. da Silva, and R. Guerra, “Preliminary results on the non-destructive determination of pear (Pyrus Communis L.) cv. Rocha ripness by Visible/Near Infrared reflectance spectroscopy,” *Acta Horticulturae*, vol. 800, pp. 1099–1106, 2008.

[25] A. Brázio, A. M. Cavaco, and R. Guerra, “A simplified two layer model for light diffuse reflectance in thin skin fruits,” *Progress in Agricultural Engineering Sciences*, vol. 6, no. 1, pp. 35–72, Dec. 2010.

[26] D. P. Ariana and R. Lu, “Evaluation of internal defect and surface color of whole pickles using hyperspectral imaging,” *Journal of Food Engineering*, vol. 96, no. 4, pp. 583–590, Feb. 2010.

[27] F. Mendoza, R. Lu, D. Ariana, H. Cen, and B. Bailey, “Integrated spectral and image analysis of hyperspectral scattering data for prediction of apple fruit firmness and soluble solids content,” *Postharvest Biology and Technology*, vol. 62, no. 2, pp. 149–160, Jul. 2011.

[28] B. M. Nicolaï, K. I. Theron, and J. Lammertyn, “Kernel PLS regression on wavelet transformed NIR spectra for prediction of sugar content of apple,” *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 2, pp. 243–252, Feb. 2007.

[29] P. Geladi and B. R. Kowalski, “Partial least-squares regression: a tutorial,” *Analytica Chimica Acta*, vol. 185, pp. 1–17, Jan. 1986.

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

[31] B. M. Wise, N. B. Gallagher, and W. Windig, *Chemometrics Tutorial for PLS _ Toolbox and Solo*. Wenatchee, WA 98801 USA: Eigenvector Research, Inc., 2006.

[32] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III, "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM Journal on Scientific and Statistical Computing*, vol. 5, no. 3, pp. 735–743, Sep. 1984.

[33] J. Kim, A. Mowat, P. Poole, and N. Kasabov, "Linear and non-linear pattern recognition models for classification of fruit from visible–near infrared spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, no. 2, pp. 201–216, Jul. 2000.

[34] Y. Liu, X. Sun, and A. Ouyang, "Nondestructive measurement of soluble solid content of navel orange fruit by visible–NIR spectrometric technique with PLSR and PCA-BPNN," *LWT - Food Science and Technology*, vol. 43, no. 4, pp. 602–607, May 2010.

[35] L. Huang, D. Wu, H. Jin, J. Zhang, Y. He, and C. Lou, "Internal quality determination of fruit with bumpy surface using visible and near infrared spectroscopy and chemometrics: A case study with mulberry fruit," *Biosystems Engineering*, vol. 109, no. 4, pp. 377–384, Aug. 2011.

[36] S. Xudong, Z. Hailiang, and L. Yande, "Nondestructive assessment of quality of Nanfeng mandarin fruit by a portable near infrared spectroscopy," *Int J Agric & Biol Eng*, vol. 2, no. 1, pp. 65–71, 2009.

[37] A. Peirs, N. Scheerlinck, B. Nicolaï Peirs, N. Temperature compensation for near infrared reflectance measurement of apple fruit soluble solids contents," *Postharvest Biology and Technology*, vol. 30, no. 3, pp. 233–248, Dec. 2003.

[38] J.-M. Roger, F. Chauchard, and V. Bellon-Maurel, "EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits," *Chemometrics and Intelligent Laboratory Systems*, vol. 66, no. 2, pp. 191–204, Jun. 2003.

[39] C. V. Greensill, P. J. Wolfs, C. H. Spiegelman, and K. B. Walsh, "Calibration Transfer between PDA-Based NIR Spectrometers in the NIR Assessment of Melon Soluble Solids Content," *Applied Spectroscopy*, vol. 55, no. 5, pp. 647–653, May 2001.

[40] P. Geladi, "Chemometrics in spectroscopy Part 1. Classical chemometrics," *Pattern Recognition*, vol. 58, no. 5, pp. 767–782, 2003.

[41] P. Geladi, B. Sethson, J. Nyström, T. Lillhonga, T. Lestander, and J. Burger, "Chemometrics in spectroscopy Part 2. Examples," *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 59, no. 9, pp. 1347–1357, Sep. 2004.

[42] A. Cavaco, "Métodos não destrutivos de análise de frutos e legumes," *Frutas legumes e flores*, pp. 63–64, 2009.

[43] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, Mar. 1993.

[44] A. Cavaco, G. Miguel, D. Antunes, and R. Guerra, "Determination of Geographical and Botanical Origin of Honey: From Sensory Evaluation to the State of the Art of Non-Invasive Technology," in *Honey: Production, Consumption and Health Benefits*, G. Bondurand and H. Bosch, Eds. Nova Science Publishers, Inc., 2011.